Stochastic Multi-host Transmission Models for Parasitic Diseases

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Dominik Heinzmann

von

Visperterminen VS

Promotionskomitee

Prof. Dr. Andrew Barbour (Vorsitz)
Prof. Dr. Paul Torgerson
Prof. Dr. Valerie Isham (London)
Prof. Dr. Leonhard Held

Zürich, 2009

Summary

The incidence of many parasitic diseases can be reduced by the introduction of appropriate public health control measures. However, such measures are costly and, particularly in poorer areas of the world, it is important to investigate how to apply the limited funds available to achieve an optimal effect. This involves a detailed understanding of the process of transmission of infection, to the point at which a mathematically based model - either analytic, or in the form of a computer simulation program - can be validated against experimental data. Given adequate models, the intervention programs can be tested and adapted if necessary.

In the present work, we focus on the transmission dynamics of particular macroparasites (Anderson & May 1982, p.300) such as helminths and arthropods. These show different biological and epidemiological behavior than microparasites (Anderson & May 1982, p.300) such as bacteria and viruses. Microparasites increase rapidly in number once introduced into a susceptible host, whereas macropoarasites in general do not reproduce, and their life-cycle is sustained by free-living larvae and in intermediate hosts. For the latter, it is important to consider the actual number of parasites in hosts, since the distribution of parasites within the hosts influences the probability of an infectious contact in either of the host populations. The model (macro)parasite in this work is *Echinococcus granulosus*, which causes a dangerous zoonotic infection. Canids are the definitive hosts of these parasites, whilst a variety of mammals act as intermediate hosts. In particular, we will focus on dogs as definitive and sheep as intermediate hosts. Humans are an ecologically aberrant host, and following infection with eggs, the larval or metacestode stage of the life cycle is pathogenic to them, and results in space occupying cystic lesions, most commonly in the liver.

Many empirical studies of *Echinococcus granulosus* show that the distribution of parasites among hosts is aggregated in the sense that only a few hosts harbor almost all parasites. As an empirical fit to such aggregated parasitic data, the negative binomial distribution is often used. Its probability mass function is $f(y;k,p) = (\Gamma(k+y)/(\Gamma(k)y!))p^k q^y$ with $k > 0$ and $0 < p = 1 - q < 1$ and Γ denotes the gamma function. The parameter k is often considered as an aggregation index since for $k \to \infty$ and $p/k \to \lambda$, the negative binomial distribution converges to the Poisson distribution with mean λ. Hence smaller values of k implies larger aggregation in the parasite distribution. In most applications, a fixed k is assumed to investigate how aggregation influences the transmission dynamics. However, a fixed k makes it difficult to investigate the (mechanistic) process which gave rise to the aggregation.

Other studies use mechanistic models to better understand the sources of aggregation in parasite loads. Most of these models describe only parts of the complete life-cycle of the parasite under investigation, or simplify the multi-host system by not explicitly modeling the intermediate host. Another problem is that for many mechanistic models, there are no appropriate data available to fit the parameters in the model.

In the present work, we introduce a mechanistic model to describe the whole life-cycle of *Echinococcus granulosus* between dogs and sheep. The model describes the evolution of the distribution of parasites in the hosts and the inter-population infections. All model parameters can be fitted to commonly available data and thus the model allows one to test many biological hypotheses, such as clumped infection and heterogeneity in susceptibility to infection. The architecture is based on two sub-processes, that model the parasite abundances in the host populations, together with a contact scheme for the inter-population infections. Compound processes are used to model the acquisition of hydatid cysts in sheep, and are fitted based on data sets from different countries. It is shown that sheep are heterogeneous with respect to infection and that they ingest clumps which are aggregated. For the dog population, shot noise processes are used to model clumped infections in dogs, where the parasite burdens in dogs decline over time. Fitting the processes to data sets from different countries indicates that the infections of dogs with *Echinococcus granulosus* occur at a low rate, but that the ingested parasite load per clump is in the thousands.

The final model links these two models by superposing a biologically reasonable infection contact pattern between the hosts, yielding a model for the whole life-cycle of the parasite. The influence of environmental factors and intervention programs on the transmission dynamics of the parasite are investigated by simulation of the final model. It is shown that the sheep population act as buffer on external perturbations of the transmission cylce.

Zusammenfassung

Kontrollmassnahmen für parasitäre Krankheiten können dazu benutzt werden, die Anzahl der Neuinfektionen zu senken und somit die Ausbreitung der Krankheit zu vermindern. Solche Interventionen sind meistens sehr kostspielig. Daher ist es vor allem in ärmeren Ländern wichtig, die limitierten finanziellen Ressourcen bestmöglichst einzusetzen um einen optimalen Effekt zu erzielen. Dies ist nur möglich, wenn der Übertragungsmechanismus der Krankheit bekannt ist. In manchen Fällen ist dieser Mechanismus sehr komplex und man benötigt analytische oder auf Simulationen basierende mathematische Modelle als Vereinfachung. Ein solches Modell kann dann mit realen Daten validiert werden und dazu verwendet werden, verschiedene Kontrollmassnahmen zu evaluieren.

In dieser Arbeit konzentrieren wir uns auf besondere Makroparasiten (Anderson & May 1982, S.300) wie zum Beispiel Würmer und Arthropoden die sich biologisch und epidemiologisch wesentlich von Mikroparasiten (Anderson & May 1982, S.300) unterscheiden. Mikroporasiten, wie zum Beispiel Bakterien und Viren, vermehren sich sich sehr schnell, sobald sie einen Wirt befallen haben. Makroparasiten dagegen reproduzieren sich im Allgemeinen nicht. Ihr Lebenszyklus wird durch frei in der Umwelt liegende Larven und durch Zwischenwirte aufrechterhalten. Es ist wichtig, die Verteilung der Anzahl Parasiten in den Wirten zu kennen, da diese einen starken Einfluss auf die Art und Anzahl von Infektionen im gesamten Übertragungszyklus ausüben. Wir fokusieren in dieser Arbeit auf den Hundebandwurm (*Echinococcus granulosus*), welcher als gefährliche Zoonose bekannt ist. Neben dem Hund als Endwirt konzentrieren wir uns auf Schafe, dem wichtigsten Zwischenwirt. Menschen können als zufällige Zwischenwirte agieren, wenn sie mit Eiern aus Hundekot in Kontakt kommen. Es bildet sich dann vor allem in der Leber Zysten, die sehr gross werden können und die Gesundheit des Menschen gefährden.

Empirische Studien des Hundebandwurms zeigen, dass die Verteilung der Anzahl Parasiten pro Wirt aggregiert ist, d.h. dass eine kleine Anzahl von Wirten den grössten Teil der Parasiten besitzen. Eine Negative Binomialverteilung wird oft benutzt um solche Daten von parasitären Krankheiten zu beschreiben. Die Wahrscheinlichkeitsmassfunktion dieser Verteilung ist $f(y;k,p) = (\Gamma(k+y)/(\Gamma(k)y!))p^k q^y$ mit $k > 0$ and $0 < p = 1 - q < 1$ und Γ bezeichnet die Gammafunktion. The Parameter k wird oft als Aggregierungsindex bezeichnet, da für $k \to \infty$ and $p/k \to \lambda$ die negative binomial Verteilung gegen die Poisson Verteilung mit Erwartungswert λ konvergiert. Kleinere Werte von k implizieren eine grössere Aggregierung in der Verteilung der Parasiten zwischen den Wirten. In vielen Applikationen nimmt man an, dass k fix ist um die Übertragungsmechanismen besser untersuchen zu können. Aber eine Fixierung von k macht es schwierig, die (mechanistischen) Prozesse zu untersuchen, welche für die Aggregierung der Parasiten verantwortlich sind.

Andere Studien benützen mechanistische Modelle, um die Ursachen der Aggregierung zu untersuchen. Die meisten solcher Modelle beschreiben aber nur einen Teil des gesamten Übertragungszyklus des untersuchten Parasiten, oder sie vereinfachen

den gesamten Zyklus indem zum Beispiel der Zwischenwirt nicht explizit modelliert wird. Ein zusätzliches Problem ist dass die Parameter vieler solcher Modelle nicht geschützt werden können, da keine passenden Daten vorhanden sind.

In dieser Arbeit führen wir ein mechanistisches Modell ein, um die Entwicklung und Ausbreitung des Hundebandwurms in Hunden und Schafen zu beschreiben. Das Modell beschreibt die Evolution der Parasiten Verteilung in den Wirten und die Infektionen zwischen den Wirten. Alle Parameter des Modelles können basierend auf vorhandenen Daten geschätzt werden, uns somit kann mit Hilfe des Modelles biologische Hypothesen getestet werden, zum Beispiel ob die Wirte mit Klumpen, welche mehrere Parasiten enthalten, infiziert werden, oder ob die Wirte heterogen sind mit Bezug auf die Anfälligkeit mit Infektionen. Die Modellarchitektur basiert erstens auf zwei Teilprozessen, welche die Infektion und ihr Verlauf in den Hunden respektive in den Schafen modelliert und zweitens auf Kontaktprozessen, welche die Infektionen zwischen den Wirten beschreiben. Compound Prozesse werden benützt, um die Entstehung von Zysten in Schafen zu modellieren. Die Prozessparameter werden basierend auf Datensätzen von verschiedenen Ländern geschützt, und es wird gezeigt, dass Schafe heterogen sind bezüglich Infektionsanfälligkeit, und dass Schafe mit Klumpen infiziert werden, welche aggregiert sind. Die Infektion mittels Klumpen in zusammen mit einer zeitlichen Abnahme der Parasiten Ladung in der Hundpopulation wird mittels Shot noise Prozessen dargestellt. Die Schätzung der Modellparameter basierend auf Datensätze aus verschiedenen Ländern impliziert, dass die Rate, mit der sich Hunde mit *Echinococcus granulosus* infizieren, klein ist, dass aber die Grösse der aufgenommenen Klumpen in der Ordnung von mehreren tausend Parasiten ist.

Das Endmodell fügt diese Teilprozesse zusammen indem es die Interaktion der Hunden mit den Schafen zusätzlich definiert, und so folglich den gesamten Übertragungszyklus des Parasiten beschreibt. Der Einfluss von Umweltfaktoren und Kontrollmassnahmen auf den Übertragungsmechanismus werden untersucht mittels Simulationen des Endmodelles. Es wird gezeigt, dass die Schafpopulation als Buffer für externe Einflüsse auf die Transmissionsdynamik dient.

Acknowledgements

Many people encouraged me along my dissertation.

First and foremost I want to thank Prof. Andrew Barbour, my main doctoral supervisor, for many fruitful discussions and suggestions and for careful reading of all of my writing. He not only served as my supervisor but also encouraged and challenged me throughout my research project. I also want to thank my second doctoral supervisor Prof. Paul Torgerson for his continuous support in particular for the biological aspects of my work and for his encouragement to participate in international research projects related to my dissertation. Together, my supervisors provided a truly interdisciplinary research atmosphere which is, in my opinion, well reflected in the dissertation.

I am grateful to many assistants from the Institute of Mathematics and the Institute of Parasitology, UZH, as well as from the Department of Mathematics, ETHZ, for many productive discussions and social activities. A special thank goes to Simon Rüegg for many philosopical discussions contributing to my understanding of the world dynamics, and all former and active members of the BAPS research group.

I also would like to thank Prof. Peter Deplazes for valuable insight into the biology of my model parasite *Echinococcus granulosus* and Carsten Rose, our system administrator, who kept the servers running for my various computations.

I enjoyed to travel to many research events all over the world, Malaysia and Kyrgyzstan in Central and Southeastern Asia, Colorado in the US, France, Germany, Italy, Irland, Spain and the UK in Europe and St. Kitts in the Caribbean. In all places, I met inspiring and motivating people who I want to acknowledge here.

Finally, the acknowledgements would not be complete without a heart felt thanks to my family and dear friends who supported me throughout my dissertation.

Dominik Heinzmann
Zürich, June 2009

Contents

Introduction

The present chapter is structured as follows. In Section 1, we briefly discuss some commonly used modeling approaches for macroparasitic diseases and the resulting ideas which led to the present work. Macroparasites (Anderson & May 1982, p.300) such as helminths and arthropods show a significant different biological and epidemiological behavior in and between hosts from that of microparasites (Anderson & May 1982, p.300) such as bacteria and viruses. Microparasites are characterized by direct multiplication within the definitive hosts and a tendency to induce immunity to reinfection in hosts. Macroparasites do in general not multiply within the definitive hosts, but produce transmission stages such as eggs and larvae which pass into the external environment. Macroparasitic infections usually depend on the number of infectious stages ingested by the hosts. This complicates the modelling of parasite dynamics, because the infection level in a population is not adequately described by the mean worm burden alone. Under natural conditions, the induced immune response in hosts normally regulates the parasite burdens in hosts, but does not protect against further infection. In what follows, the word parasite refers to macroparasites unless otherwise noted. Note that the references for this chapter are listed at the end of the dissertation.

The model parasite in the dissertation is *Echinococcus granulosus*. Here, there is extensive data available. The parasite's life cycle is introduced in Section 2, providing the basis for our mechanistic model. The model is introduced in Section 3, which gives a broad overview of the first three papers of the dissertation. Section 4 briefly discusses the fourth paper which deals with extinction times in branching processes. Finally, other modeling papers with different parasites, originated during the author's doctoral studies, are briefly discussed in Section 5. Even if the modeling approach here is tailored to the transmission dynamics of *Echinococcus granulosus*, it can be adjusted to many other parasite-host interaction systems.

1. Modeling aggregation in macroparasitic multi-host systems

Many empirical studies of *Echinococcus granulosus* (Budke et al. 2005, Gemmell et al. 1986, Torgerson et al. 2003b,c) and of many other macroparasites (Anderson 1974, Balling & Pfeiffer 1997, Boag et al. 1989a, Woolhouse et al. 1997) show that the distribution of parasites among hosts is aggregated, in the sense that only a few hosts harbor almost all parasites. The effect of aggregation can be accounted for by assuming a given form such as the commonly used negative binomial distribution for the distribution of parasites among hosts. The probability mass function of the negative binomial distribution is $f(y;k,p) = (\Gamma(k+y)/(\Gamma(k)y!))p^k q^y$ with $k > 0$ and $0 < p = 1 - q < 1$ and Γ denotes the gamma function, so that the mean of the distribution is kp and the variance is $kp(1+p)$. The parameter k is often considered

as an aggregation index since for $k \to \infty$ and $p/k \to \lambda$, the negative binomial distribution converges to the Poisson distribution with mean λ. Hence smaller values of k implies larger aggregation in the parasite distribution. The negative binomial distribution has been used in Balling & Pfeiffer (1997) to model the abundance of the fluke *Diplostomum spathaceum* in fish, in Bliss & Fisher (1953) for *European red mite* on apple leaves, in Budke et al. (2005) for the tapeworms *Echinococcus granulosus and multilocularis* in dogs, in Tanner et al. (1980) for the nematode *Trichinella spiralis* in rabbits and in Zhang et al. (2008) for the larval stage of the mites *Allothrombium pulvinum Ewing* in lice. Theses applications show that the negative binomial distribution is useful for summarizing a set of observations with two parameters. However, several quite different mechanisms can generate data which conform to the negative binomial, so that it is difficult to justify a particular mechanism merely from observing a negative binomial distribution in the data (Bliss & Fisher 1953). In addition, estimation of the dispersion parameter k for data sets with only a few positive counts (common for parasitic disease data) is quite unstable (Lloyd-Smith 2007) and thus k can strongly vary between different samples involving the same parasite, complicating the analysis and interpretation of the data and comparison of the results between the studies. One possibility is to fix k while investigating how aggregation influences the transmission dynamics. However, a fixed k makes it difficult to investigate the (mechanistic) process which gave rise to the aggregation (Pugliese et al. 1998).

Alternatively, mechanistic models can be used to better understand the mechanisms leading to aggregation in the parasite distribution in hosts. A vital source of such aggregation is infection of hosts by parasite clumps rather than by single parasite ingestions (Herbert & Isham 2000, Luchsinger 2001, Tallis & Leyton 1969). The size of the clumps may substantially vary, thus increasing the variability of the parasite distribution in hosts. Additional sources of aggregation are heterogeneity in exposure to infection of hosts or heterogeneity in the immune response of hosts.

Roberts et al. (1986) introduced an abundance-based model to describe the *Echinococcus granulosus* life-cycle. The model consisted of integrodifferential equations for the mean number of worms in dogs and cysts in sheep. Animals were assumed to lose infection independently of their parasite burden. The model mechanistically described the prevalence and the development of the mean burden of parasites in the host, but not that of the parasite densities. To describe the aggregation in the data, they fitted a negative binomial distribution, but without making a link to the ages of the animals. Barbour & Kafetzaki (1991) and Luchsinger (2001) used infinite compartmentalisation of hosts, according to their burdens of $0, 1, 2, 3, \ldots$ parasites per host, to model the transmission of *Schistosomiasis* between the definitive host, humans, and the intermediate host, water snails. The approach is based on clumped infection and provides a mechanistic description of the observed aggregation of parasites in humans. The intermediate host is not explicitly modelled. They assume that there is no superinfection in humans, so that humans acquire all their

burden of infection at once, and then enjoy complete concomitant immunity to further infections until the current infection has been eliminated. Pugliese et al. (1998) modelled the transmission dynamics between hosts and free-living larvae with a infinite system of differential equations based on clumped infections. They assumed then that parasites are distributed in hosts according to a negative binomial distribution, leading to a simplified four-dimensional system. They discussed qualitatively the behavior of the system. However, for a parasitic disease which can be described by this model, the estimation of the model parameters, such as for example the rate at which larvae are produced by adult parasites, is difficult, since appropriate data sets are in general not available. Herbert & Isham (2000) used a model that allows for several parasite stages, clumped infections and between-host heterogeneity, to describe macroparasitic transmissions involving a free-living parasite stage. Intermediate hosts were not explicitly modelled. As before, estimation of the parameters is difficult since this requires the knowledge of the distribution of the numbers of parasite larvae and mature parasites in hosts and of the lifetime distribution since maturation.

There are many other mechanistic models with characteristics similar to those discussed. Even if these models provide an understanding of the mechanisms leading to aggregated parasite distributions in the hosts, they are in general challenging to fit to real data, either because of the high-dimensionality of the parameter space or because no appropriate data is available. Thus hypotheses such as clumping of infections or heterogeneity in exposure to infection of hosts are difficult to test.

The present work proposes a mechanistic individual-based simulation model for the two-host life cycle of *Echinococcus granulosus* described in Section 2, assuming clumped infections in the hosts. The model architecture consists of two stochastic sub-processes describing the infection dynamics in the definitive and intermediate host population respectively, and a contact scheme specifying the inter-population infections. Compound Poisson and shot noise processes are used as models for the infection dynamics in the host populations, to account for clumped infections. The dynamics of the intermediate host is modelled explicitly. All model parameters have a clear biological interpretation and are estimated by the maximum likelihood method based on field data, so that the resulting estimates can be compared to experimental data and data from other field studies. Different biological hypotheses such as clumped infections and heterogeneity in exposure to infection are tested. The model is introduced in Section 3 and described in detail in the first three papers of the present dissertation.

2. Life-cycle of *Echinococcus granulosus*

The macroparasitic organisms helminths are grouped into nematoda (roundworms), flukes (flatworms) and cestoda (tapeworms), all of them highly evolved metazoa with a rather complex life-cycle (Eckert et al. 2005). Helminths do not multiply in the host, but produce offspring that must exit the host to maintain the transmission.

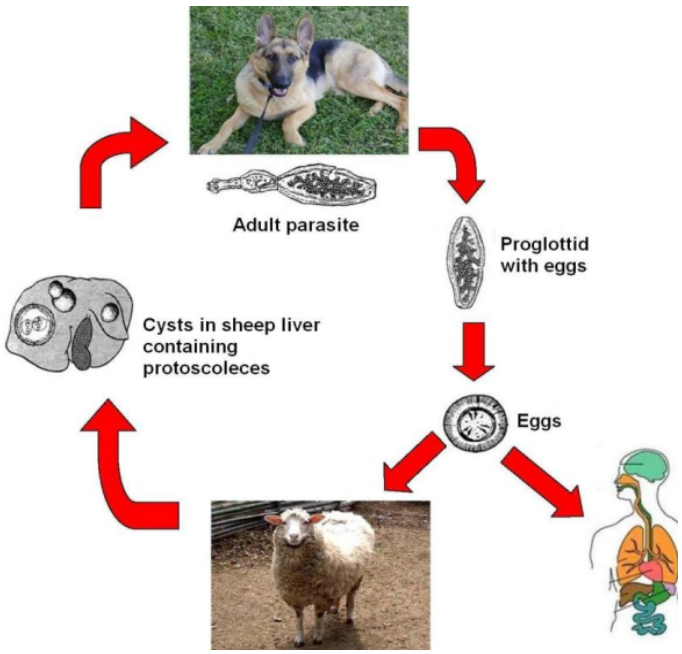

Figure 1: *Life cycle of Echinococcus granulosus. (Source: Some elements are adapted from Figure 1 in Permin & Hansen (1994).)*

The life cycle is illustrated in Figure 1, with dogs and sheep as primary definitive and intermediate hosts. The dog harbors the adult parasite in the small intestine. The adult parasite is approximately $3-10$mm long and has four suckers and a rostellum with hooks for attaching to the intestine wall. It releases eggs that are passed in the feces. The sheep ingests the eggs on pasture, which then release oncospheres that penetrate the intestinal wall and migrate through the circulatory system into various organs such as liver, brain and lungs (see Figure 2(d)). There, the oncospheres develop into cysts, within which protoscoleces develop. Hydatid cysts have a mean size of $4-7$cm, but can reach the size of a football. The development of such space occupying cystic lesions is known as cystic echinococcosis, a zoonotic parasitic disease. Humans are ecologically aberrant intermediate hosts who also develops cystic echinococcosis (Figure 2(a,b,c)). In most cases, the cysts can be surgically removed after diagnosis (Figure 2(e)). The definitive host acquires infection by ingesting organs containing cysts with protoscoleces. The protoscoleces then hatch in the small intestine and develop into adult worms. In general, the tapeworm infection does not significantly harm dogs. Alternative definitive hosts are wolves and dingos and alternative intermediate hosts are goats, cattle, horses, deer, kangaroos and camels.

The parasite is endemic in many parts of the world (Economides & Cristofi 2002, Torgerson et al. 2006) and continues to exert an unacceptable burden on human health, livestock production and wildlife ecology (Eckert & Deplazes 2004).

Figure 2: ((a)-(c)) *Hydatid cysts in humans,* (d) *infected sheep liver with hydatid cysts and* (e) *surgery of hydatid cysts in a person. (Sources:* (a,b,c) *Photos from different field studies in Kazakhstan,* (d,e) *Eckert & Deplazes (2004).)*

3. Individual-based model for the transmission of *Echinococcus granulosus*

Our model is constructed to reflect as far as possible the biological aspects of the transmission of the parasite *Echinococcus granulosus* between dogs and sheep. The model consists of two sub-processes describing the infection dynamics in the host populations, and a contact scheme specifying the inter-population infections. Compound processes are used as models for the infection dynamics in the two host populations to allow for clumped infections. The full model is described in three papers. Paper 1 proposes a model for the sheep population. The key aspect of this model is the life-long survival of cysts, which is well substantiated (Eckert & Deplazes 2004, Roberts et al. 1986, Torgerson et al. 2003b). Different mechanistic processes for modeling the acquisition of hydatid cysts are presented and fitted to data from Kazakhstan and Jordan. It is shown that a compound mixed Poisson process with a zero-truncated negative binomial distribution for the number of cysts established per clump ingested provide an adequate fit to the age-dependent cyst distribution in sheep. The random varaible Y_t modeling the total number of cysts established in an individual up to time

t is given by

$$Y_t = \sum_{k=1}^{N_t} V_k \quad \text{with} \quad \mathbb{P}(N_t = m) = \frac{\Gamma(\psi + m)}{\Gamma(\psi) m!} \left(\frac{1}{t\xi + 1}\right)^{\psi} \left(\frac{t\xi}{t\xi + 1}\right)^{m}, \qquad (2)$$

where V_k $(k = 1, 2, \ldots)$ are independent random variables describing the numbers of cysts acquired at a single infection, with common distribution $\mathcal{Q}$, and N_t is a mixed Poisson process that is independent of the V_k's. $\mathcal{Q}$ is assumed to be the zero-truncated version of a negative binomial distribution. Model (2) indicates that sheep are heterogeneous in their acquisition of infection. The parameter estimates imply that a sheep ingests an infectious clump roughly every 3 years, each clump leading on average to about $4 - 5$ established cysts.

Paper 2 introduces shot noise processes, with different time-dependent decay mechanisms for the ingested worm loads, to model the acquisition and loss of parasites in dogs. The random variable X_t modeling the total parasite load in a dog at time t is given by

$$X_t = \sum_{k=1}^{N_t} U_k h(t - \tau_k), \quad t \geq 0, \qquad (3)$$

where N_t is a Poisson random variable with mean βt, U_k $(k = 1, 2, \ldots)$ are independent and lognormally distributed random variables such that $\log(U_k)$ is $N(\mu, \sigma^2)$ distributed, and $h(t)$, $t \geq 0$, denotes the proportion of parasites still surviving t time units after infection and $0 < \tau_1 < \tau_2 \ldots$ are the times of the infection events. At each τ_i, the number of parasites that a dog ingests is modelled as a realization of a log-normal random variable. Between the τ_i's, the ingested loads decline according to $h(t)$. The experimental evidence that dogs lose the infection after a certain time (Aminzhanov 1975, Eckert & Deplazes 2004, Gemmell et al. 1986) is reflected in the different choices of $h(t)$. The models are fitted to data sets from Kazakhstan, Tunisia and China, and the estimates are shown to be plausible. Simulation studies support the good performance of the models. The results suggest that the infection rate is about 0.4, 0.6 and 0.2 infections per dog per year in Kazakhstan, Tunisia and China respectively, with corresponding means of 9000, 3000 and 1000 parasites per infection. Hence infections of dogs with *Echinococcus granulosus* occur at a low rate, but the ingested parasite load per clump is in the thousands. The mean duration of a single infection is about 8 months, comparable in all three samples.

Finally, paper 3 proposes a between population infection contact model, leading to an integrated model for the whole life cycle of *Echinococcus granulosus*. A dog may be infected when a sheep which harbors cysts containing protoscoleces dies. Since (2) only models the number of cysts in sheep, we need an additional model to describe whether cysts in sheep are fertile in the sense that they contain protoscoleces.

Let $k(t)$ be the probability that a cyst at age t has formed protoscoleces and thus is fertile. Assume that fertility is persistent, i.e. once a cyst is fertile, it remains so.

A reasonable choice for a flexible fit of $k(t)$ is the Bass model (Bass 1969) given by

$$k(t) = k^* \frac{1 - e^{-(a+b)t}}{1 + \frac{b}{a} e^{-(a+b)t}}, \tag{4}$$

where k^* is the asymptotic probability of fertility and a and b are adjustable coefficients. Our data contain records of the number of fertile and non-fertile cysts for a sheep at age t, but not the age of the cyst itself. Thus (4) is a latent process and needs to be coupled to the underlying mixed Poisson infection process of (2) in order to compare it to the data. Each animal has a fixed infection rate, and the acquisition process of cysts is Poisson. Let $q(t)$ denote the probability that a cyst is fertile in a sheep at age t. Thus the following equation for $q(t)$ is appropriate to model the fertility in cysts:

$$\begin{aligned} q(t) &= \int_0^t k(s) \frac{1}{t} ds = \frac{k^*}{t} \int_0^t \frac{1 - e^{-(a+b)s}}{1 + \frac{b}{a} e^{-(a+b)s}} ds \\ &= \frac{k^*}{t} \left[t + \frac{1}{b} \left\{ \log(1 + \frac{b}{a} e^{-(a+b)t}) - \log(1 + \frac{b}{a}) \right\} \right]. \end{aligned} \tag{5}$$

Model (5) can now be fitted to the data set from Kyrgyzstan. The parameter estimates indicate that cysts at age 2 have an average probability of 6% of being fertile and the asymptotic probability of fertility of a cyst as age increases is 10%, indicating that 1 out of 10 older cysts are fertile.

We can now use (5) to link the infection dynamics for sheep into dogs. The observed overall fertility of cysts in the data is about 5%, and the empirical data suggests that the positive sheep burdens are heavily skewed with most of the burdens in the range of $1 - 10$ cysts (Gemmell et al. 1986, Torgerson et al. 2003b). Thus it is likely that only a single cyst contains protoscoleces in an infective sheep. In addition, the empirical distribution of the positive protoscolex counts in cysts in the data implies that the log-normally distributed clump size in the dog model (3) is reasonable. Thus the infection of a dog is modelled as follows. Given that a sheep with n_c cysts dies at age t, a dog get infected with probability $1 - (1 - q(t))^{n_c}$ with a number of parasites governed by the law of U_k in (3).

To link the infection dynamics from dogs into sheep, we assume that the infection pressure on sheep is proportional to the prevalence of infection in dogs, with contacts occuring as a Poisson process. Hence we have now an integrated mechanistic model for the complete life-cycle of *Echinococcus granulosus*. The influence of environmental factors and intervention programs on the transmission dynamics of the parasite can then be investigated.

4. Extinction times in multitype branching processes

Paper 4 of this work is concerned with approximating the time to extinction in a Markov branching process. First, the initial model concerned with the transmission

dynamics of *Echinococcus granulosus* between dogs and sheep which motivated the paper is given. The model is coupled to a multitype Markov branching process. In the paper itself, we derive an approximation for the time of extinction in a sub-critical multitype Markov branching processes. The argument is based on the classical exponential approximation to the extinction probabilities (Athreya & Ney 1972, Harris 1963, Jagers 1975, Jagers et al. 2007, Sewastjanow 1974). These approximations are then combined with the branching property to derive a Gumbel approximation. It is shown that the bound on the error in total variation distance is inversely proportional to a positive power of a weighted sum of the number of individuals of the different types. The power depends on the means and higher moments of the offspring distribution. The accuracy of the approximation is illustrated by a model of parasitic resistance to the parasite *Toxoplasma gondii*, a serious public health problem.

As an annex to the paper, the approach is applied to the initial transmission model for *Echinococcus granulosus* and it is shown that the approximation performs well.

5. Additional publications

There are four further published papers that originated during the period of the dissertation and which are briefly mentioned here.

In Heinzmann & Torgerson (2008), different prevalence based models are introduced and compared for the transmission dynamics of *Echinococcus granulosus*. An extension of the prevalence-based model introduced in Roberts et al. (1986) allows one to test for a decreasing infection pressure with age of the dogs and thus for acquired immunity to infection against the parasite. It is shown that the infection rate for dogs is homogeneous and thus that acquired immunity in dogs is unlikely.

In Rüegg et al. (2008), a mathematical model is presented, that describes the transmission dynamics through ticks of the protozoa *Babesia caballi* and *Theileria equi* in horses by simultaneously using antigen and antibody information. Antigen information is obtained by polymerase chain reactions (PCR) and antibody information by immunofluorescence antibody tests (IFAT). Figure 3 shows one of the compartment-based models used in that paper. Different hypotheses within the model framework can be tested such as the hypothesis $f = 0$, indicating that maternal antibodies are fully protective. The subdivision of susceptibles into S_1 and S_2 allows the testing of an age-dependent infection rate. The model provides a biologically meaningful description of the underlying transmission process. The maximum likelihood parameter estimates resulting from fitting the model to serological data of domestic horses from Mongolia were in line with experimental data. In particular, it was shown that combining antigen and antibody information is beneficial by comparing the model to a purely antigen information based model. It was also shown that the transmission dynamics of *Babesia caballi* and *Theileria equi* are significantly different, and that maternal antibodies for *Babesia caballi* are protective against infection.

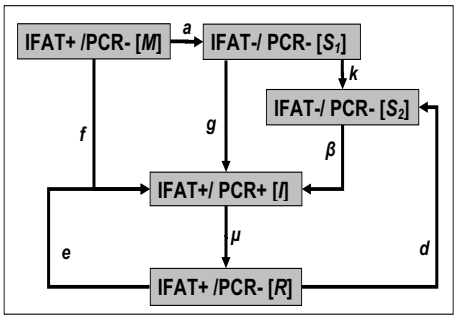

Figure 3: *Compartment model used in Rüegg et al. (2008) to test if antigen and antibody informations combined allow a more accurate description of the infection dynamics of the protozoa Babesia caballi and Theileria equi. The compartments are: Animals having maternal antibodies (M), young (S_1) and old (S_2) susceptible animals, infected animals (I) and animals eliminated the parasite (R); with corresponding rates. PCR+/- and IFAT+/- denote presence/absence of the parasite or antibodies.*

In Flütsch et al. (2008), a case-control study on farm-level was set up to identify risk factors for bovine cysticercosis in Switzerland, caused by the helminth *Taenia saginata*. We identified the following factors as being positively associated with the occurrence of bovine cysticercosis: the presence of a railway line close to cattle feeding areas, leisure activities around these areas, use of purchased roughage and organised public activities on farms attracting visitors. This information is useful for the authorities when implementing control strategies, as well as for farmers who wish to take measures tailored to their local situations.

Finally in Rapsch et al. (2008), an interactive map is created in order to demonstrate the risk of transmission, by modeling the environmental conditions that promote the survival and reproduction of the larval stages of the trematode *Fasciola hepatica* and its intermediate host, snails. The underlying model evaluates a monthly infection risk for ruminants based on measures of temperature, rainfall, soil conditions including ground water and forest cover, on a 100×100m grid over Switzerland. This was the best possible resolution compatible with the available data. The risk is categorized into levels relative to the highest value obtained over the year. The interactive map enables the user to evaluate the risk for each month and each cell of the grid over Switzerland. The model predicts that the highest transmission risk is around October which is in line with field data. The map can be used as a support for control programs for the parasite and it helps farmers to plan the pasture of their animals.

Compound processes as models for clumped parasite data

Dominik Heinzmann[1,2], A.D. Barbour[1], and Paul R. Torgerson[2,3]
[1]Institute of Mathematics, University of Zurich, Switzerland
[2]Institute of Parasitology, University of Zurich, Switzerland
[3]School of Veterinary Medicine, Ross University, West Indies

Abstract

Compound processes are proposed as models for the acquisition of hydatid cysts in sheep, caused by the parasite *Echinococcus granulosus*. The hypothesis of a clumped infection process against single ingestions is tested and it is shown that the clump-based approach provides a more accurate description of the two data sets investigated. Models with simple and mixed Poisson incidence processes and different clump size distributions are compared. A mixed Poisson incidence process with a zero-truncated negative binomial distribution for the clump sizes is shown to give an adequate description, suggesting that the acquisition of hydatid cysts in the sheep population is heterogeneous, and that the clump sizes are aggregated. The estimates of the parameters derived from the data take plausible values. The average infection rate and the clump size distribution are comparable in both data sets. Goodness-of-fit measures indicate that the model fits the data reasonably well.

Keywords: Compound processes, clumped infection, mixed Poisson, parasite data, *Echinococcus*.

1. Introduction

Parasitic disease data often consist of counts of a parasite (or an intermediate stage) in an animal, together with the animal's age. The data typically exhibit two well-known features, a substantial proportion of zeros and skewed positive counts [1, 2, 3], meaning that some hosts harbor many parasites while most have just a few. To analyze such aggregated parasite data, the fitting of the negative binomial distribution is a common method, as in [4] to model the abundance of the fluke *Diplostomum spathaceum* in fish, in [5] for *European red mite* on apple leaves, in [6] for the tapeworms *Echinococcus granulosus and multilocularis* in dogs, in [7] for the nematode *Trichinella spiralis* in rabbits and in [8] for the larval stage of the mites *Allothrombium pulvinum Ewing* in lice. However, these models do not take into account the

age of the hosts, which is known to influence the parasite pattern [9, 10, 11]. To incorporate age, negative binomial regression can be used, as in modeling the age-dependent frequency of the nematode *Wuchereria bancrofti* in humans [12], or of the nematodes *Ostertagia gruehneri* and *Marshallagia marshalli* in reindeer [13]. The approaches in both studies allow one to model (exponentially) increasing or decreasing mean parasite burdens as a function of age, in the latter study with a rather complicated relation between the over-dispersion parameter and mean of the negative binomial distribution and the covariate age. However, they do not provide any biological reason as to why this should occur.

While the negative binomial model takes aggregation into account, it may not adequately deal with high numbers of parasite-free hosts. For that purpose, zero-inflated (ZI) models [14, 15, 16] and two-part conditional (TPC) models [17, 18] can be used. These have been shown to outperform the negative binomial regression [19] for applications with an excess of zeros. These models introduce a state A in which the only counts are zeros, and a state B, in which the counts could be either zeros or positive values (ZI), or only positive values (TPC). The model parameters are p_A, the probability to be in state A, and the parameters of the conditional distribution given state B. The parameters (or combinations thereof) can be allowed to depend on covariates. In [20], a ZI negative binomial regression was applied to model egg counts of different gastrointestinal nematodes in fecal samples from young cattle by parametrizing p_A and the mean of the negative binomial distribution as functions of age. A TPC was used in [21] for modeling the density of the nematode *Wuchereria bancrofti* in mosquitoes. They argued that a zero count of microfilariae in the blood sampled by a mosquito can arise either because the human bitten is uninfected or because the blood taken from an infected human happened to contain no microfilariae. They fitted a negative binomial TPC to the aggregated data, but did not attempt to fit the underlying age-dependent model that they envisaged, because of its prohibitive complexity.

Alternatively, mechanistic models are used to understand the mechanisms leading to aggregation in the parasite distribution in hosts. A vital source of such aggregation is the infection of hosts by parasite clumps rather than single parasite ingestions [22, 23]. [24] and [25] used infinite compartmentalisation of hosts, according to their burdens of $0, 1, 2, 3, \ldots$ parasites per host, to model the transmission of *Schistosomiasis* between the definitive hosts, humans, and the intermediate hosts, water snails, by assuming clumped infections. The intermediate host is not explicitly modelled and they assume that there is no superinfection in humans. [26] used moment closure equations to describe the immuno-epidemiology of trochostrongylid nematodes in wild ruminant populations. The infection of hosts is modelled by an (inhomogeneous) compound Poisson process to account for clumped infections, and they consider nonlinear effects such as immunity and parasite-induced host mortality. Their model contains many parameters; some were fixed based on values from other studies, and the remainder were estimated from the model. However, their

model describes the mean parasite burden, but not the prevalence of infection in animals. [27] modelled the transmission dynamics between hosts and free-living larvae with a infinite system of differential equations based on clumped infections, allowing for superinfection. Then they assumed that parasites are distributed in hosts according to a negative binomial distribution, leading to a simplified four-dimensional system, whose qualitative behavior they discussed. However, it is difficult to estimate the model parameters for such diseases, as for example the rate at which larvae are produced by adult parasites, since appropriate data sets are in general not available. [22] used a model that allows several parasite stages, clumped infections and between-host heterogeneity, to describe macroparasitic transmissions involving a free-living parasite stage. As before, estimation of the parameters is difficult since this requires the knowledge of the distribution of the numbers of parasite larvae and mature parasites in hosts and of the life distribution since maturation.

In this paper, biologically interpretable mechanistic models for hydatid cysts in sheep, caused by the parasite *Echinococcus granulosus* (E.g.) [1, 28], are discussed. E.g. causes echinococcosis, a (re-)emerging hydatid disease in many parts of the world and, in particular, in Eastern Europe and the former Soviet Union [29, 30, 31]. E.g. is also potentially dangerous for humans. For this disease, it can be assumed that the cysts survive their hosts but do not replicate, and that there is no parasite-induced mortality and no acquired immunity in sheep [1, 32]. This implies a simpler infection dynamics than for example that encountered by [22] and [26]. Compound processes ([33, p.49], [34, p.25], [35, p.22]) are used to investigate the biological hypotheses that clumped (super)infections and heterogeneity in the acquisition of infection in the host population can explain the substantial proportion of zeros and thus the prevalence of infection, and the skewed positive counts of E.g. cysts in sheep.

The processes explicitly describe the underlying infection process and thus allow a natural modeling of aggregation and excess of zeros of the parasite distribution in the hosts. The prevalence and intensity is described simultaneously. The parameters can be estimated based on (standard) field data containing age and cyst counts of sheep. Goodness-of-fit measures are introduced to assess the performance of the model.

Based on two data sets from Kazakhstan [3] and Jordan [2], it is shown that clumped acquisition of infection by biologically heterogeneous hosts, where the clump sizes are aggregated, provides a satisfactory fit. Heterogeneity of acquisition of clumped infections may result from behavioral differences of sheep on pasture, or from differences in the immune system of sheep. Aggregation of clump sizes are reasonable given the highly aggregated adult parasite distribution in the definitive host, the dog [1]. Fitting the models yields parameter estimates which take biologically reasonable values. Goodness-of-fit measures indicate the reasonable performance of the model.

2. Data sets and models

2.1. Empirical data

The data sets used in this paper are from Kazakhstan [3] and Jordan [2]. The Kazakhstan sample contains 2505 individual reports of the variables age and hydatid cyst burden in sheep, caused by the parasite *E. granulosus* (E.g.) [32]. The Jordan sample counts 832 individual reports of the same variables.

Hydatid cysts develop conditional on ingestion of infective biomass by sheep (intermediate host) from contaminated environment. Contamination is caused by dogs (definitive host), which harbor adult E.g. worms in the intestine and release infective eggs in the feces. Hydatid cysts form in organs such as the liver ($60-70\%$), lungs and brain and develop over a period of years in the sheep. Cysts do not proliferate inside their hosts, but protoscoleces are produced inside the cysts which play a role in the infection of the definitive host [32]. It can be assumed that cysts survive their hosts, that there is no parasite-induced mortality and no acquired immunity in sheep [1, 32].

The records were obtained at necropsy in abattoirs with examination of the viscera of the sheep, including the lungs and liver, for the presence of hydatid cysts. The ages of the sheep were estimated from the stage of dentition and by questioning the owners of the animals. Small immature cysts were not recorded, as resources were not available for the systematic slicing of organs. A more detailed discussion of the applied sampling frame can be found in [2] and [3].

In the Kazakhstan sample, the mean and median ages are 2.037 and 2 years respectively. The interquartile range is $1-3$ years and the maximum age is 8 years. The prevalence in sheep is $0.363\ (0.344, 0.382)$. Conditional on infection, a proportion of $0.774\ (0.745, 0.800)$ harbors $1-10$ cysts, $0.186\ (0.161, 0.213)$ $11-30$ cysts and the remaining $0.041\ (0.029, 0.056)$ have more than 30 cysts. The maximal cyst burden is 64. In the Jordan sample, the mean and median ages are 2.267 years and 1 year respectively. The interquartile range is $0.5-4$ years and the maximum age is 10 years. The prevalence is $0.293\ (0.263, 0.325)$. Conditional on infection, a proportion of $0.672\ (0.609, 0.730)$ have $1-10$ cysts, $0.234\ (0.183, 0.293)$ $11-30$ and $0.094\ (0.062, 0.140)$ harbor more than 30 cysts. The maximal burden is 80 cysts. The observations in both samples agree with other study areas in Central Asia [31].

2.2. Compound Poisson process

The positive cyst burdens of *E. granulosus* in sheep are in general in the range of $1-80$ cysts per sheep [1, 3, 36]; the majority of cyst counts in sheep in both our data sets are rather low, with a large proportion of zeros. Since there is no acquired immunity in hosts [37, 38], and cysts survive for the lifetime of the sheep, the observations suggest a low infection rate and clumped ingestions of infective eggs. Sheep potentially make many random contacts with infective dog feces on

pasture, but only a small proportion of the contacts lead to an infection. Thus the resulting infection process can be viewed as a thinning of the point process at which contacts with potential infective dog feces are made. A reasonable assumption for E.g. is that the transmission system of the parasite is in a steady state [2, 28, 36], so that the ingested clumps can be supposed to be identically distributed and the low incidence rate can be supposed to be constant. Additionally, we assume that clumps are independent since infected dogs spread their feces widely, so that consecutive infections of a sheep are likely to be due to feces from different dogs. Possible clustering due to reinfection of a sheep with the same feces can be neglected since clumps in the environment have a relatively short survival time and the incidence rate is low.

The above assumptions make compound processes [33, 34, 35] a suitable choice for modeling the cyst burdens in sheep. Let the random variable Y_t denote the total number of cysts established in an individual up to time t. Then

$$Y_t = \sum_{j=1}^{N_t} S_j \,,$$

where $(N_t)_{t\geq 0}$ is a Poisson process with constant rate μ describing the number of clumps ingested by an individual sheep during the time interval $[0, t]$ and S_j $(j = 1, 2, \ldots)$ are i.i.d. random variables with distribution $\mathcal{Q}$ on the positive integers $\mathbb{N}$, independent of N_t, which describe the numbers of successfully established cysts per ingested clump. The distribution of Y_t is given by

$$\mathcal{P}_t = \sum_{k=0}^{\infty} \mathbb{P}(N_t = k)\mathcal{Q}^{*k} = \sum_{k=0}^{\infty} \frac{e^{-\mu t}(\mu t)^k}{k!}\mathcal{Q}^{*k} \,, \tag{1}$$

where $\mathcal{Q}^{*k}$ is the kth convolution of $\mathcal{Q}$. In particular,

$$p_0(t) := \mathbb{P}(Y_t = 0) = e^{-\mu t} \,. \tag{2}$$

The expectation and the variance of Y_t are

$$\mathbb{E}(Y_t) = \mathbb{E}(N_t)\mathbb{E}(S_1) \quad \text{and} \quad \mathrm{Var}(Y_t) = \mathbb{E}(N_t)(\mathrm{Var}(S_1) + [\mathbb{E}(S_1)]^2) \,.$$

2.3. Compound mixed Poisson process

To account for possible heterogeneity in the rate of acquisition of clumped infections within the sheep population, for example caused by differential immune response between sheep, the Poisson process $(N_t)_{t\geq 0}$ with fixed rate μ can be replaced by a mixed Poisson process $(\tilde{N}_t)_{t\geq 0}$, where the infection rate is a nonnegative random variable M.

It follows that

$$\mathbb{P}(\tilde{N}_t = n) = \int_0^{\infty} \frac{e^{-\mu t}(\mu t)^n}{n!} dH(\mu) \,, \tag{3}$$

where $H(\mu) = \mathbb{P}(M \leq \mu)$ and $H(0) = 0$. The distribution function H of M is also referred to as the structure distribution of the mixed Poisson process [39]. A special case is the simple Poisson process where the random variable M is degenerate at some $\mu > 0$. Mixed Poisson processes are particular examples of Cox processes or doubly stochastic Poisson processes [35, p.7].

An appropriate choice of H in (3) should provide a reasonably close approximation to the true distribution, should be easy to fit and should yield a useful interpretation of the parameters. The two-parameter gamma distributions offer a flexible and tractable family, with parameters conveniently identified as measures of skewness and scale. Let H be the distribution function of a gamma distributed random variable with shape and scale parameters $\psi, \xi > 0$ such that

$$dH(\mu) = \frac{1}{\xi^{\psi}\Gamma(\psi)}\mu^{\psi-1}e^{-\frac{\mu}{\xi}}d\mu\,, \tag{4}$$

where Γ is the gamma function. Then

$$\mathbb{P}(\tilde{N}_t = n) = \frac{t^n}{\xi^{\psi}\Gamma(\psi)n!}\int_0^{\infty}\mu^{\psi+n-1}e^{-\mu\frac{t\xi+1}{\xi}}d\mu$$

and, since $\int_0^{\infty} z^n e^{-az}dz = n!a^{-n-1}$,

$$\mathbb{P}(\tilde{N}_t = n) = \frac{\Gamma(\psi+n)}{\Gamma(\psi)n!}\left(\frac{1}{t\xi+1}\right)^{\psi}\left(\frac{t\xi}{t\xi+1}\right)^{n}. \tag{5}$$

Equation (5) describes a negative binomial distribution, with $\mathrm{Var}(\tilde{N}_t) > \mathbb{E}(\tilde{N}_t)$, where

$$\mathbb{E}(\tilde{N}_t) = \psi\xi t =: at \quad \text{and} \quad \mathrm{Var}(\tilde{N}_t) = (\psi\xi t)(1+\xi t) =: at + bt^2\,. \tag{6}$$

Using (5) in (1), the distribution of Y_t becomes

$$\tilde{\mathcal{P}}_t = \sum_{k=0}^{\infty}\frac{\Gamma(\psi+k)}{\Gamma(\psi)k!}\left(\frac{1}{t\xi+1}\right)^{\psi}\left(\frac{t\xi}{t\xi+1}\right)^{k}\mathcal{Q}^{*k}\,. \tag{7}$$

In particular,

$$\tilde{p}_0(t) := \tilde{\mathbb{P}}(Y_t = 0) = \left(\frac{1}{t\xi+1}\right)^{\psi}, \tag{8}$$

where $\tilde{\mathbb{P}}$ is the probability measure under $\tilde{N}_t$ as counting process. Setting $\xi = \mu/\psi$, for fixed n, t and μ, (5) becomes

$$\begin{aligned}\mathbb{P}(\tilde{N}_t = n) &= \mu\frac{\psi+n-1}{t\mu+\psi}\mu\frac{\psi+n-2}{t\mu+\psi}\cdots\mu\frac{\psi}{t\mu+\psi}\left(\frac{\mu t}{\psi}+1\right)^{-\psi}\frac{t^n}{n!}\\ &\xrightarrow{\psi\to\infty}\frac{e^{-\mu t}(\mu t)^n}{n!}\,,\end{aligned} \tag{9}$$

where the exponential term in the limit is based on Euler's formula $\exp(x) = \lim_{N\to\infty}(1+(x/N))^N$, for any real x. The limit is thus a Poisson distribution.

3. Decompounding and estimation

Decompounding [40] defines the procedure of obtaining the base distribution $\mathcal{Q}$ and the Poisson rate parameter μ based on a sample of the compound process $(\mathcal{P}_t)_{t\geq 0}$. Given a parametric form of the discrete distribution $\mathcal{Q}$, the convolution $\mathcal{Q}^{*k}$ can easily be computed and (1) respectively (7) can be fitted to the data by the maximum likelihood estimation method. This approach is easy to implement and provides reasonable computational performance, since cyst burdens in sheep are mostly rather low, the maximal burdens being of magnitude 80. Since $\mathcal{Q}$ is defined on the positive integers, $\mathcal{Q}^{*k}$ needs only be computed for small k's. In addition, simulation from the fitted model is computationally fast (we will use the fitted model in a subsequent paper).

A nonparametric alternative to estimate the distribution $\mathcal{Q}$ is presented in [40]. Using an empirical estimator for the distribution of Y_t for t fixed, an estimator for the distribution of the S_i's is obtained by a suitable inversion of the Panjer recursions [41] of the distribution of Y_t. As shown in [40], the procedure requires an accurate empirical estimation of the distribution of Y_t for each t. Since the sheep in our sample are of many different ages and the loads are heavily skewed, it is difficult to obtain an appropriate empirical estimate of the distribution of Y_t for the nonparametric procedure.

Suppose that $\mathcal{Q}$ is the zero-truncated Po(η) distribution. Then the following result [42] is useful.

Theorem 3.1. *Let S_j $(1 \leq j \leq n)$ be i.i.d. zero-truncated Po(η) distributed random variables, so that $\mathbb{P}(S_j = s) = \eta^s/(s!(e^\eta - 1))$ for $s \in \mathbb{N}$. Then for $z \in \mathbb{N}$,*

$$\mathbb{P}(\sum_{j=1}^{n} S_j = z) = \begin{cases} \frac{\eta^z}{z!(e^\eta-1)^n} \sum_{k=0}^{n} (-1)^k (n-k)^z \binom{n}{k} & \text{if } n \leq z \\ 0 & \text{else.} \end{cases}$$

To take into account aggregation of the clump size distribution, let $\mathcal{Q}$ be the zero-truncated negative binomial distribution, so that for $s \in \mathbb{N}$,

$$\mathbb{P}(S_j = s) = \frac{\Gamma(\theta + s)}{\Gamma(\theta) s!} \frac{\left(\frac{\zeta}{\zeta+1}\right)^y}{(1+\zeta)^\theta - 1}, \tag{10}$$

where θ is the shape and ζ is the scale parameter of the negative binomial distribution. Then the following results [43] applies.

Theorem 3.2. *Let S_j $(1 \leq j \leq n)$ be i.i.d. zero-truncated negative binomial distributed random variables specified by (10). Then for $z \in \mathbb{N}$,*

$$\mathbb{P}(\sum_{j=1}^{n} S_j = z) = \begin{cases} \frac{\left(\frac{\zeta}{\zeta+1}\right)^z \left(\frac{1}{\zeta+1}\right)^{\theta n}}{\left[1-\left(\frac{1}{\zeta+1}\right)^\theta\right]^n} \sum_{k=1}^{n} (-1)^{n-k} \binom{n}{k} \binom{\theta k + z - 1}{z} & \text{if } n \leq z \\ 0 & \text{else.} \end{cases}$$

Let $\mathbb{P}_\Omega$ be the probability measure corresponding to the compound Poisson process if $\Omega = \mu$ and to the compound mixed Poisson process if $\Omega = (\psi, \xi)$; let N_t denote the corresponding incidence process. Then, $\mathbb{E}(Y_t|N_t = n) = n\mathbb{E}(S_1)$ and $\mathrm{Var}(Y_t|N_t = n) = n\mathrm{Var}(S_1)$. Hence for the a zero-truncated Poisson clump distribution,

$$\mathbb{E}(Y_t|N_t = n) = \frac{n\eta}{1 - e^{-\eta}} \ , \quad \mathrm{Var}(Y_t|N_t = n) = \frac{n\eta}{1 - e^{-\eta}} \left(1 - \frac{\eta}{e^\eta - 1}\right) ,$$

and for a zero-truncated negative binomial clump distribution,

$$\mathbb{E}(Y_t|N_t = n) = \frac{n\theta\zeta}{1 - (1/(\zeta + 1))^\theta} \tag{11}$$

and

$$\mathrm{Var}(Y_t|N_t = n) = n \left[\frac{\theta\zeta(1 + \zeta + \theta\zeta)}{1 - (1/(\zeta + 1))^\theta} - \left(\frac{\theta\zeta}{1 - (1/(\zeta + 1))^\theta} \right)^2 \right] . \tag{12}$$

Expressions (1) and (7) can be used with Theorems 3.1 and 3.2 to compute the unconditional distribution of Y_t,

$$\mathbb{P}_\Omega(Y_t = j) = \begin{cases} \mathbb{P}_\Omega(N_t = 0) & \text{if } j = 0 \\ \sum_{k=1}^{j} \mathbb{P}_\Omega(N_t = k)\mathbb{P}(\sum_{l=1}^{k} S_l = j) & \text{if } j \geq 1 \, . \end{cases} \tag{13}$$

Given independent realizations y_i $(1 \leq i \leq n)$ of Y_t at time points t_i, the log-likelihood function is

$$\begin{aligned} l(\Omega, \eta) = \sum_{i=1}^{n} \Big\{ & I_{\{y_i=0\}} \ln \mathbb{P}_\Omega(N_t = 0) \\ & + I_{\{y_i>0\}} \ln \left[\sum\nolimits_{k=1}^{y_i} \mathbb{P}_\Omega(N_t = k)\mathbb{P}(\sum\nolimits_{l=1}^{k} S_l = y_i) \right] \Big\} , \end{aligned} \tag{14}$$

where I is the indicator function. The log-likelihood function for the case of a single ingestion mechanism, with clump size fixed to be 1, is thus

$$l_2(\Omega) = \sum_{i=1}^{n} \ln \mathbb{P}_\Omega(N_t = y_i) \, . \tag{15}$$

Let us introduce the following model notation for the rest of the paper. The single ingestion models with Poisson and mixed Poisson incidence process are denoted by P/1 and MP/1 respectively. The compound process $(Y_t)_{t\geq 0}$ (13) with $(N_t)_{t\geq 0}$ a Poisson process and with the clump size distribution $\mathcal{Q}$ specified to be the zero-truncated Poisson distribution is denoted by P/ztP, and if $(N_t)_{t\geq 0}$ is a mixed Poisson process, then the model is denoted by MP/ztP. Analogously, if the clump size distribution is specified to be the zero-truncated negative binomial distribution, we denote the resulting models by P/ztnb and MP/ztnb, depending on the incidence process.

4. Application

Parameter estimates for the models of interest are obtained from the two data sets of Kazakhstan and Jordan (Section 2.1). We test single against clumped infection, heterogeneity of the Poisson rate parameter of the incidence process, and aggregation of the clump size distribution. Then we compare the best fitting models for the two data sets and assess the goodness-of-fit.

4.1. Clumped infection

First, we compare the single ingestion models P/1 and MP/1 to the compound processes P/ztP and MP/ztP respectively using a standard likelihood ratio test based on (14) and (15) with 1 degree of freedom. The log-likelihood values are reported in Table 1. Testing the P/1 against the P/ztP results in p-values of < 0.001 for Kazakhstan and Jordan. Similarly, testing the MP/1 against the MP/ztP also results in p-values of < 0.001 for Kazakhstan and Jordan. Hence there is strong evidence for a clumped infection process in both samples.

Table 1: Log-likelihood values for the models fitted to the Kazakhstan and Jordan samples, together with the number of parameters in the models.

Model	Kazakhstan	Jordan	Parameters
P/1	−10648.570	−2643.109	1
MP/1	−4230.321	−1133.255	2
P/ztP	−4647.557	−1161.142	2
P/ztnb	−4179.769	−1018.524	3
MP/ztP	−4180.413	−1079.412	3
MP/ztnb	−4160.347	−1016.665	4

4.2. Heterogeneity in acquisition and aggregated clump sizes

In (9), we have seen that, if $\xi = \mu/\psi$ with μ fixed and $\psi \to \infty$, then the MP/ztP model converges to the P/ztP model. To test if the acquisition of hydatid cysts of sheep is heterogeneous, we have to test the null hypothesis $H_0 : \xi = 0$ against $\xi > 0$. Analogously, to test if the clump size distribution is aggregated, we note that if $\zeta = \eta/\theta$ with η fixed and $\theta \to \infty$, then the P/ztnb model converges to the P/ztP model, and thus we need to test $H_0 : \zeta = 0$ against $\zeta > 0$. Clearly, the MP/ztP and the P/ztnb models are also nested within the MP/ztnb model, which allows heterogeneity in the acquisition of cysts together with an aggregated clump size distribution. For the tests with $H_0 : \xi = 0$ and $H_0 : \zeta = 0$, we test a parameter which is on the boundary of the parameter space under H_0. [44] showed that the asymptotic distribution of the likelihood ratio test statistic in the presence

of a parameter that is on the boundary of the null hypothesis is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, a $50:50$ mixture of χ_0^2 and χ_1^2 distributions. Given the observed test statistic $\bar{\chi}$, the p-value is given by $(\mathbb{P}(\chi_0^2 > \bar{\chi}) + \mathbb{P}(\chi_1^2 > \bar{\chi}))/2$.

Applying the likelihood ratio test with the above asymptotic χ^2 mixture distribution to the reported log-likelihood values in Table 1 implies that the P/ztnb and the MP/ztP model both fit the Kazakhstan and Jordan sample significantly better than the P/ztP (all p-values smaller than 0.001). In addition, the MP/ztnb fits the two samples significantly better than the P/ztnb (p-values for Kazakhstan < 0.001 and Jordan 0.027) and the MP/ztP models (p-values for Kazakhstan and Jordan < 0.001).

To verify the asymptotic distribution of the test statistic under H_0, we apply a Monte Carlo method and simulate data under H_0 (simpler model), then fit both the simpler and more complex model to the generated data sets and compute the likelihood test statistic. For the generation of the data sets, starting with the original ages t_k $(1 \leq k \leq n)$ of the n sheep in the sample, a new cyst burden is attributed to each of them as a realization of the simpler model with $t = t_k$, with the model parameters fixed at their estimated values given in Table 2. Repeating this procedure 2000 times yields an approximating reference distribution of the test statistic under H_0. Testing the P/ztnb model against the MP/ztnb model for the Jordan sample implies a p-value of 0.035, which is slightly larger than the p-value of 0.027 obtained by using the asymptotic reference distribution. The other p-values computed with the simulated reference distribution also differ slightly from the ones obtained with the asymptotic reference distribution, however they are also smaller or equal to 0.002. It appears that our samples are too small to be able to rely completely on asymptotics. However, the test results with the simulated reference distribution also imply that the MP/ztnb model significantly better fits the data sets from Kazakhstan and Jordan than the other models. We conclude that there is evidence in the data that the acquisition of hydatid cysts of *Echinococcus granulosus* by sheep is heterogeneous, and that the clump size distribution is aggregated.

Table 2 shows the estimates of the MP/ztnb model for the parameters $a = \psi\xi$ and $b = \psi\xi^2$ of the incidence process N_t defined in (6) and for the mean $c := \mathbb{E}(Y_t|N_t = 1)$ and variance $d := \mathrm{Var}(Y_t|N_t = 1)$ of the clump size distribution defined in (11) and (12). The parameter a is not significantly different in the samples from Kazakhstan and Jordan, suggesting that a sheep gets infected on average every third year. The parameter b is significant larger in the Kazakhstan sample, so that the variance of the infection rates $\mathrm{Var}(N_t) = at + bt^2$ is larger for this sample. The difference of the variance of the infection rate in the two samples is especially pronounced in older sheep since $\mathrm{Var}(N_t) \sim bt^2$. The resulting gamma mixture distributions (4) of the infection rate for the two samples are plotted in Figure 1, indicating that in the Kazakhstan sample, the infection rates are more heterogeneous than in the Jordan sample. Table 2 also indicates that the estimated mean and variance for the clump size distribution are not significantly different in the two samples, suggesting that the

Table 2: Maximum likelihood estimates for the parameters and key quantities of the MP/ztnb model for the Kazakhstan and Jordan samples, together with 95% confidence intervals computed by the bootstrap percentile method. The parameters $a = \psi\xi$ and $b = \psi\xi^2$ of the incidence process N_t are defined in (6), so that $\mathbb{E}(N_t) = at$ and $\mathrm{Var}(N_t) = at + bt^2$. The mean $c := \mathbb{E}(Y_t|N_t = 1)$ and variance $d := \mathrm{Var}(Y_t|N_t = 1)$ of the clump size distribution are defined in (11) and (12) respectively.

	Kazakhstan	Jordan
$\hat{\psi}$	0.941 (0.629, 1.260)	5.154 (2.579, 8.061)
$\hat{\xi}$	0.343 (0.225, 0.741)	0.060 (0.029, 0.173)
$\hat{\theta}$	0.351 (0.139, 0.617)	0.212 (0.126, 0.442)
$\hat{\zeta}$	5.859 (3.215, 9.763)	7.861 (5.394, 10.565)
$\hat{a}$	0.323 (0.237, 0.499)	0.309 (0.195, 0.521)
$\hat{b}$	0.111 (0.064, 0.198)	0.019 (0.008, 0.061)
$\hat{c}$	4.186 (2.343, 6.276)	4.500 (2.724, 7.022)
$\hat{d}$	19.798 (10.177, 29.828)	27.125 (14.411, 35.917)

number of successfully established cysts per infection is similar in the two samples. Thus on average, an infective clump leads to about $4 - 5$ established cysts in the sheep.

The fitted MP/ztnb model provides estimates for the prevalence of infection as well as for the probability mass function (pmf) of the positive loads. Figure 2 shows the estimated prevalence of infection for the MP/ztnb model for the Kazakhstan and Jordan samples together with the observed prevalences. In both samples, the estimated prevalence of the MP/ztnb explains the observations reasonably well.

The estimated pmf of the MP/ztnb model for the age classes reported in Figure 2 are displayed in Figure 3 for the Kazakhstan and in Figure 4 for the Jordan sample. Given an age class, the fitted pmf are computed as mixture of the pmf's corresponding to the different ages within the class. The fitted pmf are reasonable in both samples, taking into account the small number of observed positive loads in some of the age classes, especially in the Jordan sample.

4.3. Goodness-of-fit

The goodness-of-fit of the MP/ztnb model is evaluated as follows. Divide the sheep into age classes, and treat the observations in the different classes as i.i.d. data. The classes are specified as in Figure 2. The observed and estimated distributions of cysts are then compared within each age class using an appropriate statistic. Note that, as before, the resulting pmf for an age class is a mixture of the pmf's corresponding to the different ages within that class.

With the age classes as before, let n_i $1 \leq i \leq 6$ be the number of animals in

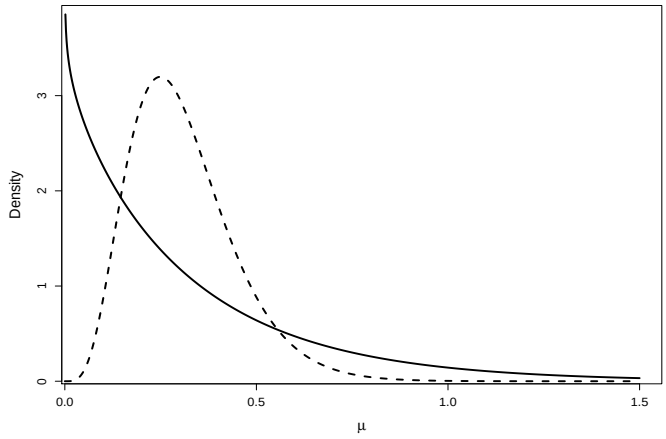

Figure 1: Estimated gamma density function (4) of the infection rate in the incidence process for the samples from Kazakhstan (solid line) and Jordan (dashed line).

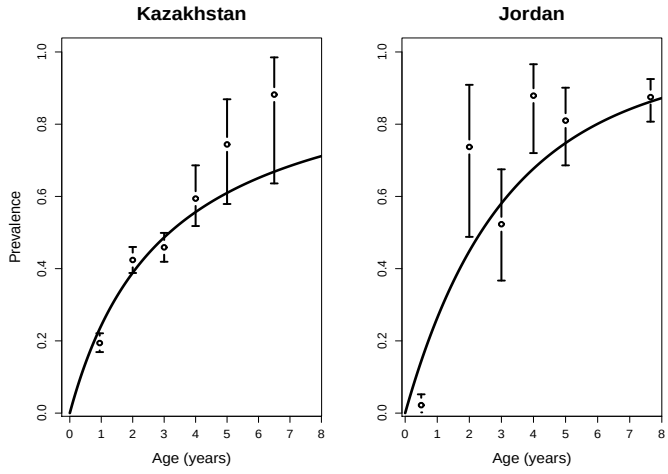

Figure 2: Fitted prevalence curves $\hat{q}(t) = 1-1/(t\hat{\xi}+1)^{\hat{\psi}}$ (with $\hat{\psi}$ and $\hat{\xi}$ given in Table 2) for the MP/ztnb model for the samples from Kazakhstan and Jordan, together with the observed prevalences and their 95% confidence intervals. The observed prevalences are computed for the age classes $(0,1]$, $(1,2]$, $(2,3]$, $(3,4]$, $(4,5]$, 5+, where 5+ summarizes all sheep older than 5 years. For the age classes $1-4$, the majority of the observed ages coincide with the end points of the interval. The prevalences are plotted at the means of the ages of the animals in the corresponding classes.

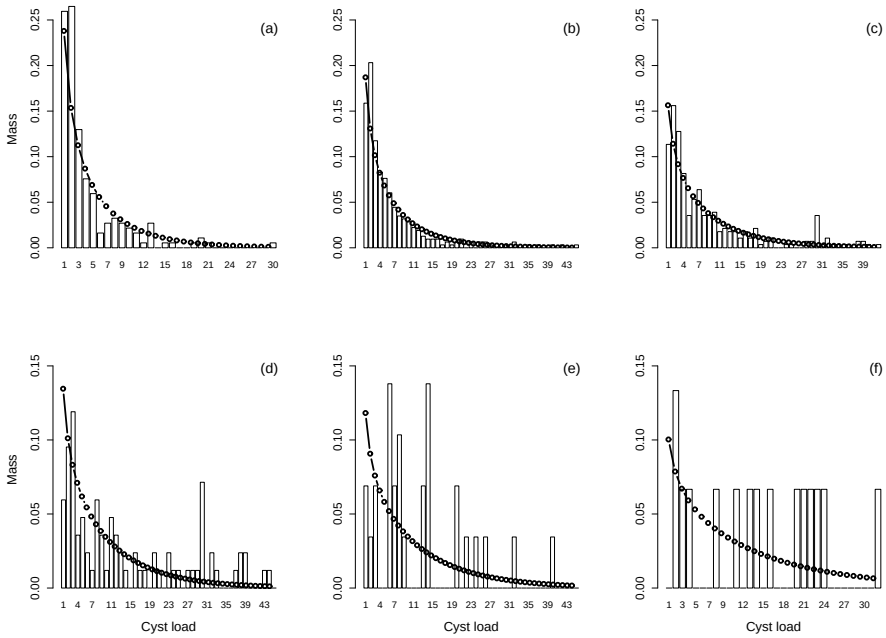

Figure 3: Estimated probability mass functions of the MP/ztnb model for the positive loads of the Kazakhstan sample for the age classes (a) $(0, 1]$, (b) $(1, 2]$, (c) $(2, 3]$, (d) $(3, 4]$, (e) $(4, 5]$, (f) 5+, together with a histogram of the corresponding observed quantities. The class sizes are 185, 315, 282, 84, 29 and 15. For a better presentation of the results, the following points are not plotted in the histograms: 64 (with corresponding mass 0.003) in age class $(1, 2]$, 47 and 57 (mass 0.119 each) in age class $(3, 4]$ and 56 (mass 0.034) in age class $(4, 5]$.

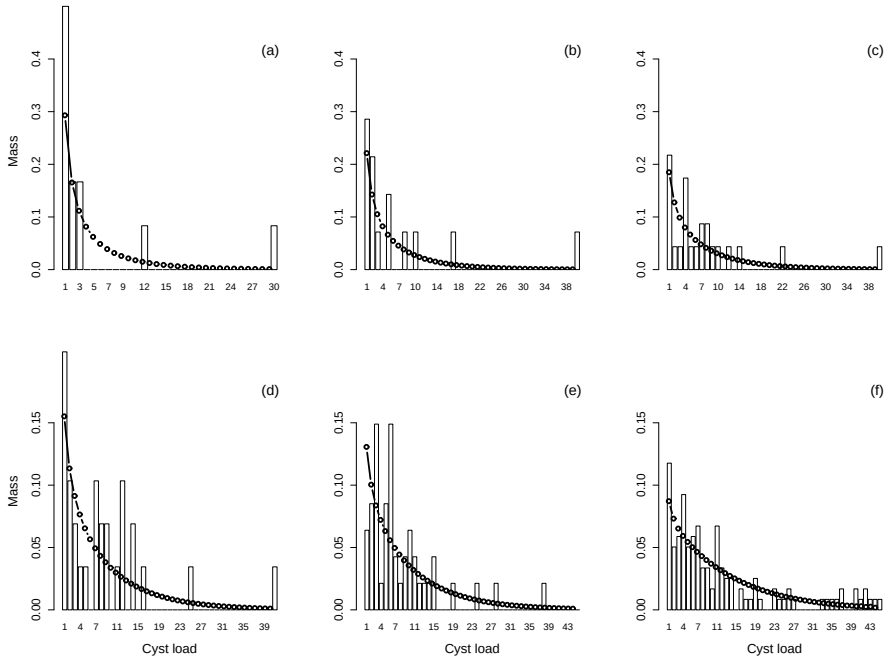

Figure 4: Estimated probability mass functions of the MP/ztnb model for the positive loads of the Jordan sample for the age classes (a) $(0, 1]$, (b) $(1, 2]$, (c) $(2, 3]$, (d) $(3, 4]$, (e) $(4, 5]$, (f) 5+, together with a histogram of the corresponding observed quantities. The class sizes are 12, 14, 23, 29, 47 and 119. For an better presentation of the results, the following points are not plotted in the histograms: 52 and 63 (mass 0.021 each) in age class $(4, 5]$, and 57 (mass 0.008) and two loads of 80 (combined mass 0.016) in age class 5+.

age class i, and stratify them with respect to load into c_i strata. Then two possible goodness-of-fit measures for the distribution of the numbers of cysts within any given age class i are

$$\chi^2 := \sum_{k=1}^{c_i} \frac{(m_{ik} - \mathbb{E}(M_{ik}))^2}{\mathbb{E}(M_{ik})}$$

and

$$L := \sum_{k=1}^{c_i} \left| \mathbb{E}\left(\frac{M_{ik}}{n_i}\right) - \frac{m_{ik}}{n_i} \right| ,$$

where M_{ik} is a random variable describing the numbers of animals of age class i having cyst counts in stratum k ($1 \leq k \leq c_i$), and m_{ik} is the (corresponding) observed count.

The number of strata c_i for age class i is chosen to be the maximal number such that the expected number of counts in each stratum is at least 10. The strata in the age classes are computed for the model with parameters fixed by their estimates in Table 2. To generate the reference distribution of χ^2 and L, a Monte Carlo approach is used, where data sets are generated under the MP/ztnb model. Given the original ages t_k ($1 \leq k \leq n$) of the n sheep in the sample, a new cyst burden is attributed to each of them as a realization of the MP/ztnb model with $t = t_k$ and the parameters fixed by their estimates given in Table 2. We then fit the MP/ztnb model to this new data set, and compute with the new estimates the test statistics for each of these sets. We use the same stratification of the age classes as before. The observed values of the two test statistics can then be compared to the reference distributions for each age class i.

Figures 5 and 6 display the results for the samples from Kazakhstan and Jordan for 1000 simulations. For the Kazakhstan sample, the observed values of the test statistics χ^2 and L (indicated by a solid line) although consistently large, are in reasonable agreement with the simulated distributions for all age strata. For the Jordan sample, the solid line lies well outside the simulated distribution in age class $(0, 1]$. This is for two reasons. First, the observed prevalence in that age class is overestimated by the model (see Figure 2). Secondly, there are only 12 positive loads in that class, which are not well described by the model. However, the results in the other age classes suggest that the model fit is reasonable.

The model seems to have some tendency to underestimate the zero load stratum and to overestimate the numbers of high cyst counts in the first age class. The opposite tendency can be observed in the age strata $4 - 6$. Since only 4 parameters are used in the model, to fit the distributions of prevalence and cyst burden observed in 6 different age classes, a perfect fit can hardly be expected.

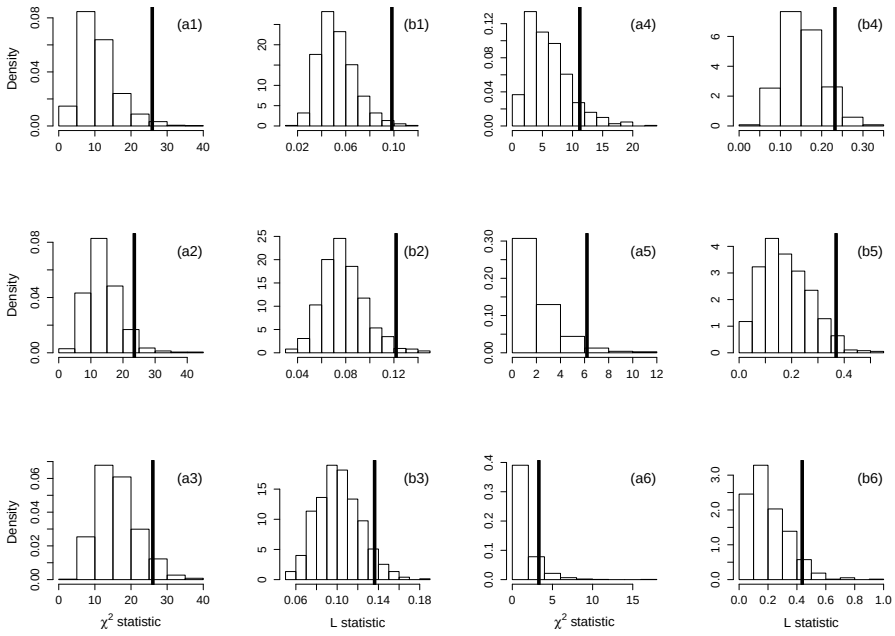

Figure 5: Goodness-of-fit of the MP/ztnb model in the Kazakhstan sample. The observed values of the test statistics (solid lines) χ^2 ((a1)-(a8)) and L ((b1)-(b8)) are plotted with the corresponding simulated distributions under the MP/ztnb model with parameters fixed with its estimates in Table 2 for the age classes (x1) $(0, 1]$, (x2) $(1, 2]$, (x3) $(2, 3]$ (x4) $(3, 4]$, (x5) $(4, 5]$ and (x6) 5+, with x=a,b.

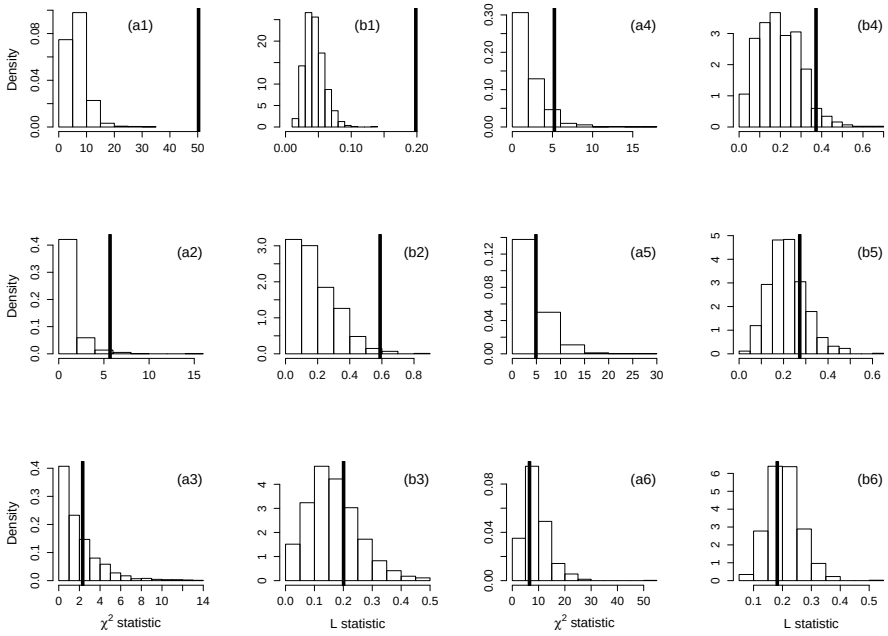

Figure 6: Goodness-of-fit of the MP/ztnb model in the Jordan sample, analogously to Figure 5.

5. Conclusion

In this paper, different mechanistic models are used to explain the acquisition of hydatid cysts in sheep, caused by the parasite *Echinococcus granulosus*. The models allow one to test the biologically interesting hypotheses of clumped infections, host heterogeneity with respect to infection and aggregation of clump sizes. The experimentally supported assumptions of *Echinococcus granulosus* cysts infections in sheep such as life-long survival of cysts in the host, no replication inside the host, no parasite-induced mortality and no acquired immunity in sheep imply simpler infection dynamics than for example those encountered by [22] and [26], as discussed in the introduction to this paper. Hence our models are straightforward to fit to the most commonly available data sets, which only contain the ages and cyst burdens of the sheep. The models provide age-dependent estimates for the prevalence of infection and for the probability mass functions of positive cyst burdens in sheep.

The application of the models to two data sets from Kazakhstan and Jordan supports a clumped infection process, with a rate of acquisition of infection which is heterogeneous within the population, and with clump sizes which are aggregated. The infection process is described by a compound mixed Poisson process with a zero-truncated negative binomial distribution for the number of cysts per ingested clump. The goodness-of-fit measures indicate that the chosen model describes the given data reasonably well, but not perfectly. The estimates suggest a mean infection rate of about 0.315 infections per year and a mean clump size of about 4.5 cysts, suggesting that on average every third year, a sheep will ingest an infectious clump, each clump leading to approximately $4 - 5$ established hydatid cysts in the sheep. The results indicate that the observed aggregation in the distribution of cysts among sheep may be the result both of differences between sheep and also of clumped infections.

Our model can be used to investigate how changes in the underlying parameters may affect the parasite distribution, and thus may be useful in assessing control programs for *Echinococcus granulosus*. In particular, it can be used as sub-process for describing infections in the sheep population in a fully stochastic model for the complete life-cycle of *Echinococcus granulosus*.

Achnowledgements The authors gratefully acknowledge the comments and suggestions of two referees and the handling editor, that greatly improved the presentation. This work was supported by the Schweizerischer Nationalfonds (SNF), project no. 107726.

References

[1] M. A. Gemmell, J. R. Lawson, M. G. Roberts, Population dynamics in echinococcosis and cysticercosis: biological parameters of Echinococcus granulosus in dogs and sheep, Parasitology 92 (1986) 599–620.

[2] P. R. Torgerson, D. H. Williams, M. N. Abo-Shehada, Modelling the prevalence of Echinococcus and Taenia species in small ruminants of different ages in northern Jordan, Vet Parasitol 79 (1998) 35–51.

[3] P. R. Torgerson, B. S. Shaikenov, A. T. Rysmukhambetova, A. E. Ussenbayev, A. M. Abdybekova, K. K. Burtisurnov, Modelling the transmission dynamics of Echinococcus granulosus in sheep and cattle in Kazakhstan, Vet Parasitol 114 (2003b) 143–153.

[4] T. E. Balling, W. Pfeiffer, Frequency distributions of fish parasites in the perch *Perca fluviatilis* l. from Lake Constancee, Parasitol Res 83 (1997) 370–373.

[5] C. I. Bliss, R. A. Fisher, Fitting the negative binomial distribution to biological data, Biometrics 9 (1953) 176–200.

[6] C. M. Budke, J. Qiu, P. S. Craig, P. R. Torgerson, Modeling the transmission of Echinococcus granulosus and Echinococcus multilocularis in dogs for a high endemic region of the Tibetan plateau, Int J Parasitol. 35 (2005) 163–170.

[7] C. E. Tanner, M. A. Curtis, T. D. Sole, G. K., The nonrandom, negative binomial distribution of experimental trichinellosis in rabbits, Parasitology 66 (1980) 802–805.

[8] Z. Q. Zhang, P. R. Chen, K. Wang, X. Y. Wang, Overdispersion of Allothrombium pulvinum larvae (Acari: Trombidiidae) parasitic on Aphis gossypii (Homoptera: Aphididae) in cotton fields, Ecol Entomology 18 (2008) 379–384.

[9] B. Boag, P. B. Topham, R. Webster, Spatial distribution on pasture of infective larvae of the gastro-intestinal nematode parasites of sheep, Int J Parasitol. 19 (1989a) 681–685.

[10] B. Boag, H. H. Kolb, Influence of host age and sex on nematode populations in the wild rabbit (Oryctolagus cuniculus L.), P. Helm. Soc. Wash. 56 (1989b) 116–119.

[11] S. W. Pacala, A. P. Dobson, The relation between the number of parasites per host and host age: population dynamic causes and maximum-likelihood estimation, Parasitology 96 (1988) 197–210.

[12] C. Braga, R. Ximenes, J. Miranda, N. Alexander, Bancroftian filariasis in an endemic area of Brazil: differences between genders during puberty, Rev. Soc. Bras. Med. Trop. 38 (2005) 224–228.

[13] R. J. Irvine, A. Stien, J. F. Dallas, O. Halvorsen, R. Langvatn, S. D. Albon, Life-history strategies and population dynamics of abomasal nematodes in Svalbard reindeer (Rangifer tarandus platyrhynchus), Parasitol 120 (2000) 297–311.

[14] E. Dietz, D. Boehning, On estimation of the Poisson parameter in zero-modified Poisson models, Comput. Stat. Data Anal. 34 (2000) 441–459.

[15] N. L. Johnson, S. Kotz, Distributions in Statistics: Discrete Distributions, Boston: Houghton Mifflin, 1969.

[16] C. Li, J. Lu, J. Park, K. Kim, P. A. Brinkley, J. P. Peterson, Multivariate zero-inflated Poisson models and their applications, Technometrics 41 (1999) 29–38.

[17] A. C. Cohen, An extension of a truncated Poisson distribution, Biometrics 16 (1960) 447–450.

[18] N. Duan, W. G. J. Manning, C. Morris, J. Newhouse, Choosing between the sample selection model and the multi-part model, JBES 2 (1984) 283–289.

[19] C. E. Rose, S. W. Martin, K. A. Wannemuehler, B. D. Plikaytis, On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data, J Biopharm Stat. 16 (2006) 463–481.

[20] A. Nodtvedt, I. Dohoo, J. Sanchez, G. Conboy, L. DesCôteaux, G. Keefe, L. K., J. Campell, The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows, Can J Vet Res. 66 (2002) 249–257.

[21] P. K. Das, S. Subramanian, A. Manoharan, K. D. Ramaiah, P. Vanamail, B. T. Grenfell, D. A. P. Bundy, E. Michael, Frequency distribution of Wuchereria bancrofti infection in the vector host in relation to human host: evidence for density dependence, Acta Tropica 60 (1995) 159–165.

[22] J. Herbert, V. Isham, Stochastic host-parasite interaction models, J. Math. Biol. 40 (2000) 343–371.

[23] G. M. Tallis, M. Leyton, Stochastic models of populations of helminthic parasites in the definitive host, Math Biosci 4 (1969) 39–48.

[24] A. D. Barbour, M. Kafetzaki, Modeling the overdispersion of parasite loads, Math Biosci 107 (1991) 249–253.

[25] C. J. Luchsinger, Stochastic models of a parasitic infection, exhibiting three basic reproduction ratios, J Math Biol 42 (2001)(6) 532–554.

[26] B. T. Grenfell, K. Wilson, V. S. Isham, H. E. G. Boyd, K. Dietz, Modelling patterns of parasite aggregation in natural populations: trichostrongylid nematode-ruminant interactions as a case study, Parasitology 111(Suppl.) (1995) 135–151.

[27] A. Pugliese, R. Rosa, M. L. Damaggio, Analysis of a model for macroparasitic infection with variable aggregation and clumped infections, J Math Biol 36 (1998) 419–447.

[28] R. C. A. Thompson, A. J. Lymbery, The biology of Echinococcus and hydatid disease, London: George Allen and Unwin, 1986.

[29] B. Todorov, V. Boeva, Human echinococcosis in Bulgaria: a comparative epidemiologiocal analysis, Bulletin WHO 77 (1999) 110–118.

[30] P. R. Torgerson, B. Shaikenov, K. K. Baitursinov, A. M. Abdybekova, The emerging epidemic of echinococcosis in Kazakhstan, Trans R Soc Trop Med Hyg 96 (2002) 124–128.

[31] P. R. Torgerson, B. Oguljahan, M. E. Muminov, R. R. Karaeva, O. T. Kuttubaev, M. Aminjanov, B. Shaikenov, Present situation of cystic echinococcosis in Central Asia, Parasitol Int. 55 (2006) 207–212.

[32] J. Eckert, P. Deplazes, Biological, epidemiological and clinical aspects of Echinococcosis, a zoonosis of increasing concern, Clin Microbiol Rev. 17 (2004) 107–135.

[33] D. R. Cox, V. Isham, Point Processes, New York: Chapman and Hall, 2 edition, 1980.

[34] D. J. Daley, D. Vere-Jones, An introduction to the theory of point processes, New York: Springer, 2 edition, 1988.

[35] A. F. Karr, Point processes and their statistical inference, Marcel Dekker, Inc., 2 edition, 1991.

[36] M. A. Gemmell, Hydatid disease in Australia, III. Observations on the incidence and geographical distribution of hydatidiasis in sheep in New South Wales, Aust Vet J 34 (1958) 269–280.

[37] M. G. Roberts, J. R. Lawson, M. A. Gemmell, Population dynamics in echinococcosis and cysticercosis: Mathematical model of the life-cycle of Echninococcus granulosus, Parasitology 92 (1986) 621–641.

[38] P. R. Torgerson, D. D. Heath, Transmission dynamics and control options for cystic echinococcosis, Parasitology 127 (2003d) 143–158.

[39] J. L. Teugels, P. Vynckier, The structure distribution in a mixed poisson process, JAMSA 9 (1996)(4) 489–496.

[40] B. Buchmann, R. Grübel, Decompounding Poisson random sums: Recursively truncated estimates in the discrete case, Ann. Statist. 31 (2003) 1054–1074.

[41] H. R. Panjer, Recursive evaluation of a family of compound distributions, ASTIN Bulletin 12 (1981) 22–26.

[42] J. Springael, I. van Nieuwenhuyse, On the sum of independent zero-truncated Poisson random variables, Research paper UA, Faculty of Applied Economics (2006).

[43] T. Cacoullos, C. Charalambides, On minimum variance unbiased estimation for truncated binomial and negative binomial distributions, Ann Inst Stat Math. 27 (1975) 235–244.

[44] S. G. Self, K. Y. Liang, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, JASA 4 (1987) 605–610.

Shot noise processes for clumped infections with time-dependent decay dynamics

Dominik Heinzmann[1,2], A.D. Barbour[1], and Paul R. Torgerson[2,3]
[1]Institute of Mathematics, University of Zurich, Switzerland
[2]Institute of Parasitology, University of Zurich, Switzerland
[3]School of Veterinary Medicine, Ross University, West Indies

Abstract

Shot noise processes are introduced to model aggregated parasitic count data arising from clumped superinfections coupled with decay mechanism of the ingested parasite clumps. The decay patterns considered are an exponential decay of the clump magnitude coupled with an absorption process around zero, and a constant survival time of an ingested clump. Maximum likelihood is used to fit the models to three samples from Kazakhstan, Tunisia and China, containing ages and *Echinococcus granulosus* worm counts for individual dogs. The approach is based on numerically inverting the Laplace transform for the exponential decay model and on direct computation for the model with constant survival of the clumps. It is shown that the parameter estimates take plausible values and that the decay dynamics is comparable in the three samples. The mean infection period from a single ingested clump is estimated in all models, suggesting that dogs cease to be infectious after about 8 months. The models reasonably represent the prevalence of infection and the distribution of positive parasite counts in dogs. The best fitting models suggest that the infection rate is about 0.4, 0.6 and 0.2 infections per dog per year in Kazakhstan, Tunisia respectively China, with corresponding means of 9000, 3000 and 1000 parasites per infection. Hence infections of dogs with *Echinococcus granulosus* occur at a low rate, but the ingested parasite load per clump is in the thousands.

Keywords: Shot noise process, inverse Laplace transform, clumped infection, infection duration, *Echinococcus granulosus*.

1. Introduction

In macroparasitic diseases, count data on parasites in hosts is often the result of multiple clumped superinfections and a decay pattern of the established clumps (Anderson & May 1978). The number and time points of the clumped superinfections are in general not observable and thus the decay pattern is difficult to determine. The resulting parasite counts are most likely strongly aggregated, with many zero and skewed positive counts (Gemmell et al. 1986, Woolhouse et al. 1997). Analysis

of such data is most commonly done by using descriptive tools, as for example negative binomial distributions in Budke et al. (2005), Balling & Pfeiffer (1997), negative binomial regression in Braga et al. (2005) and Irvine et al. (2000) to account for the host's age, or zero-inflated models in Nodtvedt et al. (2002) and two-part conditional models in Das et al. (1995) to account for the excess of zeros. However, the parameters of these models are in general not directly linked to the underlying transmission dynamics of the parasite. Thus important biological features, such as superinfection and parasite life history, do not appear explicitly and thus are difficult to address.

Herbert & Isham (2000) used a mechanistic model based on clumped infections to describe such data. They assumed that parasites ingested by clumps evolve independently of all others, going through different parasite stages. Their model can be used to test different scenarios such as influence of life time of the parasite or parasite-induced host mortality. However, parameter estimation is difficult since common data sets of macroparasitic diseases only contain ages and parasite counts of the hosts. Grenfell et al. (1995) and Pugliese et al. (1998) also used mechanistic models based on clumped infections to describe macroparasitic data sets. Both models allow for superinfection, and incorporate nonlinear effects such as parasite-induced immunity. As above, the resulting models are difficult to fit to commonly available data sets of macroparasitic infections.

In this paper, we present shot noise processes (Cox & Isham 1980, p.135) as models of aggregated parasitic count data arising from clumped infections coupled with a death process of the parasites. We focus on the infection process of the parasite *Echinococcus granulosus* (Thompson & Lymbery 1986) in the definitive host, a dog. *Echinococcus granulosus* is potentially dangerous for humans, who act as accidental intermediate hosts, and is endemic in many parts of the world (Budke et al. 2005, Eckert & Deplazes 2004). Dogs are infected by consuming infected viscera of sheep, the typical intermediate host which harbor infectious hydatid cysts. Assuming that dogs get infected by clumps of parasites and that the clumps then decrease over time, shot noise processes present a reasonable choice of model.

A shot noise process is a superposition of independent and identically distributed shots which occur at different time points and whose effects decay over time (Lund et al. 1999) and is thus a natural extension of compound processes, to allow for some dynamics such as a decrease of the accumulated shots (Cox & Isham 1980, p.135). The shots here represent single successful infections of a dog by a clump of parasites. The numbers and time points of the infections are described by a counting process. The dynamics between the infections that we consider in this paper are (i) exponential decay with absorption in zero and (ii) a constant survival time for clumps. Since observed parasite counts in dogs are heavily skewed (Eckert & Deplazes 2004, Torgerson et al. 2003a), shots are assumed to be lognormally distributed. The models are fitted to three samples from different countries. It is shown that the estimated values are plausible with regard to other experimental data and that all

three data sets are adequately described using the models. The mean duration of a single infection is derived for the models. The results provide new insight into the transmission dynamics of *Echinococcus granulosus*.

2. Data sets and models

2.1. Empirical data

The data samples used in this paper contain ages and *Echinococcus granulosus* parasite counts of dogs from South Kazakhstan (Torgerson et al. 2003a) with a sample size of 606 dogs, from the Testour and Bouzid regions of Tunisia (Lahmar et al. 2001) with 140 dogs and from Sichuan Province, People's Republic of China (Budke et al. 2005) with 371 dogs. The life-cycle of *Echinococcus granulosus* takes place primarily between dogs as definitive hosts and sheep as intermediate hosts (Eckert & Deplazes 2004). Dogs harbor the adult parasite in the small intestine; it releases eggs that are passed in the feces. Sheep ingest the eggs on pasture, and some of them develop into hydatid cysts. Cysts become fertile if they produce protoscoleces. Humans are ecologically aberrant intermediate hosts. Dogs acquire the infection by ingesting organs from the sheep that contain fertile cysts. The protoscoleces then hatch in the small intestine and develop into adult worms. In general, the adult parasite does not proliferate within the dog, and the tapeworm infection does not cause significant harm. The parasite is endemic in many parts of the world (Economides & Cristofi 2002, Torgerson et al. 2006) and continues to exert a burden on human health, livestock production and wildlife ecology (Eckert & Deplazes 2004).

Parasite burdens of *Echinococcus granulosus* in dogs in the above samples are obtained by purging and then collecting the intestinal contents. The dog ages are derived from an interview with the owner and a personal assessment of the animals by the interviewer. In the Kazakhstan sample, most of the dogs are free of parasites (76.9%), and 86.4% of the infected dogs harbor 1000 or fewer parasites. Similar patterns are observed in the Tunisia and China samples, where 72.9% and 91.6% respectively of the dogs are parasite free, and 86.8% and 90.3% respectively of the infected animals respectively harbor 1000 or fewer parasites. The maximal parasite loads in the Kazakhstan, Tunisia and China samples are 150000, 67000 and 20000 adult parasites. The mean ages of dogs are 3.133 years for the Kazakhstan, 4.630 years for the Tunisia and 4.237 years for the China sample.

2.2. Shot noise models

A dog normally ingests only a few fertile cysts per infection, but the number of protoscoleces in such cysts is rather high. Hence a clumped infection mechanism is realistic. Furthermore, if the infection rate of dogs is low, as suggested in Gemmell (1959), Roberts et al. (1986), Torgerson et al. (2006) and Torgerson et al. (2003a), acquired immunity of dogs can reasonably be neglected (Gemmell et al. 1986) and the

infection process can be modelled by a Poisson process. A reasonable assumption for *Echinococcus granulosus* in our study regions is that the transmission system is in a steady state (Roberts et al. 1986, Eckert & Deplazes 2004, Torgerson et al. 2006), so that since infective sheep typically harbor only one fertile cyst and since the number of protoscoleces per fertile cyst is well described by a log-normal distribution (reanalysis of data from Torgerson et al. (2009)), the clump sizes can be supposed to be identically distributed. We further assume that the clump sizes of infections of a single dog are independent since they come from different sheep, and since acquired immunity in dogs can be taken to be negligible. The number of surviving parasites from a clump decreases over time (Aminzhanov 1975, Kapel et al. 2006, Thompson & Lymbery 1986), so that shot noise process is reasonable as a model for *Echinococcus granulosus* in dogs.

A shot noise process $(X_t)_{t\geq 0}$ is a continuous-time piecewise deterministic stochastic process. Events of $(X_t)_{t\geq 0}$ occur at times $0 < \tau_1 < \tau_2 \ldots$, given by the realization of a point process N_t on the nonnegative integers $\mathbb{N}_0$. Henceforth, we shall always take N_t to be a Poisson process of rate β. At each τ_i, there is a realization of a non-negative random variable U_i known as shot effect, and $X_{\tau_i} - X_{\tau_i -} = U_i \geq 0$. Between the τ_i's, $(X_t)_{t\geq 0}$ undergoes a death process determined by a non-increasing function $h(t)$. Assume that $X_0 = 0$ almost surely. Then

$$X_t = \sum_{k=1}^{N_t} U_k h(t - \tau_k)\,, \quad t \geq 0\,, \tag{1}$$

where U_k $(k = 1, 2, \ldots)$ are i.i.d. random variables with density function f_U independent of N_t, $h(t) = 0$ if $t < 0$, and $X_t = 0$ if $N_t = 0$. In this paper, the shots U_k correspond to the number of successfully established *Echinococcus granulosus* parasites per infection of a dog.

The integral formulation of (1) is

$$X_t = \int_0^t h(t - s) dU_s^*\,,$$

where $U_s^* := \sum_{k=1}^{N_s} U_k$. Since $X_t \geq 0$ almost surely, the one-sided Laplace transform

$$L_{X_t}(s) := \mathbb{E}(e^{-sX_t})$$

exists throughout $\mathrm{Re}(s) > \gamma$ for some $\gamma \leq 0$. The random variable X_t has a point mass $F_{X_t}(0)$ at zero, and a density f_{X_t} over the interval $(0, \infty)$. The point mass $F_{X_t}(0)$ is equal to $\exp(-\beta t)$ if $h(t) > 0$ for all positive t, and it is the probability that no shots arrive during the time interval $[0, t]$. The continuous portion f_{X_t} is due to the arrival of one or more shots during $[0, t]$. The function $L_{X_t}(s)$ is completely monotonic for $\mathrm{Re}(s) > \gamma$ (Widder 1946, p.161).

Assuming that all moments of X_t exist, one has for $\mathrm{Re}(s) > \gamma$

$$\mathbb{E}(X_t^j) = (-1)^j \frac{d^j L_{X_t}(s)}{ds^j}\bigg|_{s=0} . \tag{2}$$

Since $(N_t)_{t\geq 0}$ is a Poisson process, it follows that $\mathbb{E}(X_t^j) < \infty$ if $\mathbb{E}(U_k^j) < \infty$ for any $j \geq 1$, as is seen by using $\sum_{k=1}^{N_t} U_k$ as upper bound of X_t $(t \geq 0)$.

2.3. Decay pattern

Suppose that $h(t) = \exp(-\lambda t)$ in (1). Since $X_0 = 0$ almost surely, Cox & Isham (1980, p.136) have shown that

$$L_{X_t}(s) = \exp\left\{ -\beta \int_{0+}^{t} [1 - L_U(se^{-\lambda z})] dz \right\} , \tag{3}$$

where L_U is the common Laplace transform of the U_k $(k = 1, 2, \ldots)$. Using (3) in (2),

$$\mathbb{E}(X_t) = \frac{\beta \mathbb{E}(U)}{\lambda}[1 - e^{-\lambda t}] , \quad \mathrm{Var}(X_t) = \frac{\beta \mathbb{E}(U^2)}{2\lambda}[1 - e^{-2\lambda t}] , \tag{4}$$

with $\mathbb{E}(U^j)$ the jth moment of the U_k's.

There is experimental evidence that dogs eventually lose *Echinococcus granulosus* infections (Aminzhanov 1975, Eckert & Deplazes 2004, Gemmell et al. 1986). We incorporate this with two possibilities. First, define X_t^0 to be a random variable such that $X_t^0 = 0$ if $X_t \in [0, 1)$ and $X_t^0 = X_t$ else. Thus $X_t^0 \in 0 \cup [1, \infty)$ so that

$$\begin{aligned} F_{X_t^0}(0) &= F_{X_t}(0) + \int_0^1 f_{X_t}(u) du \quad \text{if } z = 0 , \\ f_{X_t^0}(z) &= f_{X_t}(z) \qquad\qquad\qquad\qquad \text{if } z \geq 1 . \end{aligned} \tag{5}$$

This model will be referred to as MA (mass accumulation) model.

The second possibility is to use the Poisson transform. Let Y_t have the mixed Poisson distribution $\mathrm{Po}(X_t)$; that is, for a given $t \geq 0$,

$$\mathbb{P}(Y_t = y) = \int_0^\infty \frac{e^{-x} x^y}{y!} f_{X_t}(x) dx , \quad y \in \mathbb{N} , \tag{6}$$

and

$$\mathbb{P}(Y_t = 0) = e^{-\beta t} + \int_0^\infty e^{-x} f_{X_t}(x) dx = \mathbb{E}(e^{-X_t}) . \tag{7}$$

Now, for $y \in \mathbb{N}$,

$$(-1)^y \frac{d^y L_{X_t}(s)}{ds^y} = \mathbb{E}(X_t^y e^{-sX_t}) = \int_0^\infty x^y e^{-xs} f_{X_t}(x) dx \,,$$

so that from (6)

$$\mathbb{P}(Y_t = y) = \frac{(-1)^y}{y!} \frac{dL^y_{X_t}(s)}{ds^y}\bigg|_{s=1} \,, \quad y \in \mathbb{N} \,, \tag{8}$$

and also $\mathbb{P}(Y_t = 0) = L_{X_t}(1)$. This model will be referred to as PT (Poisson transform) model.

Alternatively, the decay pattern is defined as $h(t) = I(t \leq t_d)$, where I is the indicator function and t_d a fixed duration, so that all shots have a constant survival time. Then

$$X_t = \sum_{k=1}^{N_t} U_k I(t - \tau_k \leq t_d) =_d \sum_{k=1}^{N_{t \wedge t_d}} U_k \,,$$

where $=_d$ indicates equal in distribution, and so

$$L_{X_t}(s) = e^{-\beta(t \wedge t_d)(1 - L_U(s))} \,, \tag{9}$$

and

$$\mathbb{E}(X_t) = \beta(t \wedge t_d)\mathbb{E}(U) \,, \quad \mathrm{Var}(X_t) = \beta(t \wedge t_d)\mathbb{E}(U^2) \,. \tag{10}$$

This model will be referred to as CS (constant survival) model.

2.4. Shot effect distribution

Since observed parasite counts in dogs are heavily skewed (Eckert & Deplazes 2004, Torgerson et al. 2003a), shots are assumed to be lognormally distributed. Let U_k $(k = 1, 2, \ldots)$ be independent and lognormally distributed random variables, so that $\log(U_k) \sim N(\mu, \sigma^2)$. Then (4) transforms into

$$\mathbb{E}(X_t) = \frac{\beta e^{\mu + \sigma^2/2}}{\lambda}[1 - e^{-\lambda t}] \,, \quad \mathrm{Var}(X_t) = \frac{\beta e^{2\sigma^2 + 2\mu}}{2\lambda}[1 - e^{-2\lambda t}] \,, \tag{11}$$

and (10) into

$$\mathbb{E}(X_t) = \beta(t \wedge t_d) e^{\mu + \sigma^2/2} \,, \quad \mathrm{Var}(X_t) = \beta(t \wedge t_d) e^{2\sigma^2 + 2\mu} \,. \tag{12}$$

The Laplace transform of the U_k's, $L_U(s)$, is needed to evaluate (3) and thus the PT model in 8. We will see later that $L_U(s)$ is also needed to compute the MA and the CS models. There is no general closed-form expression for $L_U(s)$. However, it

can be represented through a series expansion based on Gauss-Hermite integration for s real (Mehta et al. 2007), so that

$$\begin{aligned} L_U(s) = \mathbb{E}(e^{-sU}) &= \int_0^\infty \exp(-su) \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(u)-\mu)^2}{2\sigma^2}\right] du \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{\pi}} \exp[-s\exp(\sqrt{2}\sigma z + \mu)] \exp(-z^2) dz \\ &= \sum_{i=1}^N \frac{\omega_i}{\sqrt{\pi}} \exp\left[-s\exp(\sqrt{2}\sigma a_i + \mu)\right] + R_N \,, \end{aligned}$$

where N, ω_i and a_i $(1 \le i \le N)$ in the final expression are the order, weights and nodes (abscissas) of the Gauss-Hermite integration. For small N, ω_i and a_i can be found in Abramowitz & Stegun (1972, p.924). For larger N, the integration nodes a_i are found as a root of the Hermite polynomial H_N of order N (Abramowitz & Stegun 1972, p.509), and the corresponding weights ω_i are calculated by using

$$\omega_i = \frac{2^{N-1} N! \sqrt{\pi}}{N^2 [H_{N-1}(a_i)]^2} \,.$$

The error term R_N decreases with N and several upper bounds can be found (Mehta et al. 2007). However, they are difficult to compute for N large and not sharp enough for many applications (Stoer & Bulirsch 1983, p.171). Thus numerical methods are necessary to choose an appropriate N (see Section 4).

Let $A_i := \omega_i/\sqrt{\pi}$ and $B_i := \exp(\sqrt{2}\sigma a_i + \mu)$, so that $\sum_{i=1}^N A_i = 1$ (Mehta et al. 2007). An approximation to the Laplace transform of the U_k's is then given by

$$\hat{L}_U(s) = \sum_{i=1}^N A_i \exp\left[-sB_i\right] \,. \tag{13}$$

Since $\sum_{i=1}^N A_i = 1$, we can let W denote a discrete random variable taking values B_i with probabilities A_i $(1 \le i \le N)$, and (13) implies that $\hat{L}_U$ is the Laplace transform of W and that $\hat{L}_U$ is completely monotone. Hence

$$\begin{aligned} \hat{L}_{X_t}(s) &= \exp\left\{-\beta \int_0^t [1 - \hat{L}_U(se^{-\lambda z})] dz\right\} \\ &= \exp\left\{-\beta \int_0^t \left[1 - \sum_{i=1}^N A_i \exp\left(-se^{-\lambda z} B_i\right)\right] dz\right\} \,, \end{aligned} \tag{14}$$

where $h(t) = \exp(-\lambda t)$ as before. A similar argument as above for $\hat{L}_U$ shows that $\hat{L}_{X_t}(s)$ is a Laplace transform and that it is completely monotonic.

Equation (14) indicates that

$$\lim_{s\to 0}\hat{L}_{X_t}(s)=1\ ,\ \ \lim_{s\to\infty}\hat{L}_{X_t}(s)=e^{-\beta t}\,.$$

Substituting $c=sB_i\exp(-\lambda z)$ in (14),

$$\hat{L}_{X_t}(s)=\exp\left\{-\beta t+\frac{\beta}{\lambda}\sum_{i=1}^{N}A_i\int\limits_{sB_ie^{-\lambda t}}^{sB_i}\left[\frac{e^{-c}}{c}\right]dc\right\}\,,$$

where the integral is well-defined, since $sB_i\exp(-\lambda t)>0$ if $s>0$, and is bounded above by λt. Using $E_1(a)=\int_a^\infty\frac{e^{-b}}{b}db$, $a>0$,

$$\hat{L}_{X_t}(s)=\exp\left\{-\beta t+\frac{\beta}{\lambda}\sum_{i=1}^{N}A_i[E_1(sB_ie^{-\lambda t})-E_1(sB_i)]\right\}\,. \tag{15}$$

If $h(t)=I(t\le t_d)$, with t_d the constant duration of a single infection, an approximation to $L_{X_t}(s)$ is given by

$$\hat{L}_{X_t}(s)=e^{-\beta(t\wedge t_d)(1-\hat{L}_U(s))}\,.$$

For the PT model, the probabilities (8) can now be approximated by using (15). Then $dE_1(a)/da=-\exp(-a)/a$ $(a>0)$ implies that for $s>0$

$$\frac{d\hat{L}_{X_t}(s)}{ds}=\frac{\beta}{\lambda}\left\{\sum_{i=1}^{N}A_i\left[\frac{e^{-sB_i}}{s}-\frac{e^{-sB_ie^{-\lambda t}}}{s}\right]\right\}\hat{L}_{X_t}(s)$$

and $d^2\hat{L}_{X_t}(s)/ds^2$ is

$$\begin{aligned}&\frac{\beta}{\lambda}\left\{\sum_{i=1}^{N}A_i\left[\frac{B_ie^{-\lambda t}e^{-sB_ie^{-\lambda t}}-B_ie^{-sB_i}}{s}-\frac{e^{-sB_i}-e^{-sB_ie^{-\lambda t}}}{s^2}\right]\right\}\hat{L}_{X_t}(s)\\&+\left\{\frac{\beta}{\lambda}\sum_{i=1}^{N}A_i\left[\frac{e^{-sB_i}}{s}-\frac{e^{-sB_ie^{-\lambda t}}}{s}\right]\right\}^2\hat{L}_{X_t}(s)\,,\end{aligned}$$

and hence (8) yields the probabilities. Higher order analytical derivatives can be derived using the above formulas.

2.5. Mean survival times for shots

Let T be a random variable for the duration of infection for a single shot. For the MA model, $\mathbb{P}(T\ge t)=\mathbb{P}(U\ge\exp(\lambda t))=1-\Phi((\lambda t-\mu)/\sigma)$ for any t fixed, so that

$$\mathbb{E}(T)=\frac{\sigma}{\lambda}\int\limits_{-\mu/\sigma}^{\infty}[1-\Phi(y)]dy\,.$$

For the PT model, given U, we have $Y_t \sim \text{Po}(U\exp(-\lambda t))$. Hence $\mathbb{P}(T \leq t) = \mathbb{E}[\mathbb{P}(Y_t = 0|U)] = \mathbb{E}(\exp(-U\exp(-\lambda t)))$, so that given $U = u$, $T \sim V/\lambda + \log(u)/\lambda$, where V is a Gumbel random variable. It follows that

$$\mathbb{E}(T) = \frac{\mathbb{E}(V) + \mathbb{E}(\log(U))}{\lambda} = \frac{\gamma^* + \mu}{\lambda},$$

where γ^* is the Euler-Mascheroni constant.

Finally for the CS model, the survival time is constant and thus $\mathbb{E}(T) = t_d$.

3. Likelihood inference

We have seen that X_t in (1) with $h(t) = \exp(-\lambda t)$ has a point mass of size $\mathbb{P}(X_t = 0) = F_{X_t}(0) = \exp(-\beta t)$ at zero and a continuous density f_{X_t} on $(0, \infty)$. Define

$$\tilde{L}_{X_t}(s) := \int_0^\infty e^{-sx} f_{X_t}(x)dx,$$

so that $\tilde{L}_{X_t}(s) = L_{X_t}(s) - \mathbb{P}(X_t = 0)$. Then the density of X_t on $x > 0$ can be written as

$$f_{X_t}(x) := \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{sx} \tilde{L}_{X_t}(s)ds, \tag{16}$$

where γ is chosen such that the line $s = \gamma$ lies in the complex plane to the right of all singularities of $\tilde{L}_{X_t}(s)$. Here, we take $\gamma = 0$.

The contour integral in (16) cannot be evaluated analytically and thus needs to be numerically approximated. The Stehfest algorithm for numerical inversion (Stehfest 1970) can be used. Its scheme is given for $x > 0$ by

$$f_{X_t,M}(x) = \frac{\log(2)}{x} \sum_{k=1}^{2M} \zeta_k \tilde{L}_{X_t}\left(\frac{k\log(2)}{x}\right) = \frac{\log(2)}{x} \sum_{k=1}^{2M} \zeta_k L_{X_t}\left(\frac{k\log(2)}{x}\right), \tag{17}$$

since $\sum_{k=1}^{2M} \zeta_k = 0$ so that $\sum_{k=1}^{2M} \zeta_k \mathbb{P}(X_t = 0) = 0$. The coefficients ζ_k are defined in Stehfest (1970) and M is such that increasing values of M imply a more accurate inversion. Using the approximation $\hat{L}_{X_t}$ (15) in (17) implies an approximation for the density of X_t on $x > 0$, which we denote as $\hat{f}_{X_t,M}(x)$. In this case, the Stehfest algorithm is a suitable choice since it is stable for completely monotonic functions (Abate & Valko 2004).

Given observations of loads x_l and ages t_l $(1 \leq l \leq n)$, the log-likelihood function of the MA model can be approximated as

$$\sum_{l=1}^{n} \left\{ I_{\{x_l<1\}} \log\left[e^{-\beta t} + \int_{0+}^{1} \hat{f}_{X_{T_l},M}(x)dx\right] + I_{\{x_l \geq 1\}} \log[\hat{f}_{X_{t_l},M}(x_l)] \right\}. \tag{18}$$

The probabilities $\mathbb{P}(Y_t = y)$ for the Poisson transform Y_t in (6) can be computed by using (8) which requires (analytical) evaluation of higher order derivatives of $\hat{L}_{X_t}$. However, (6) implies that for sufficient large y, $\mathbb{P}(Y_t = y)$ is close to $f_{X_t}(y)$ and thus the latter can be used for the computation. Denote such an appropriate y as y_0. The determination of y_0 is discussed in Section 4. Given y_0, the log-likelihood for the PT model becomes

$$\sum_{l=1}^{n} \left\{ I_{\{x_l < y_0\}} \log[\mathbb{P}(Y_t = x_l)] + I_{\{x_l \geq y_0\}} \log[\hat{f}_{X_{t_l},M}(x_l)] \right\} . \tag{19}$$

For the CS model, $X_t =_d \sum_{k=1}^{N_{t \wedge t_d}} U_k$ and thus the likelihood can be computed based on sums of lognormals, as in Heinzmann et al. (2009) for the compound Poisson process. Even if there is no general explicit formula for the distribution of sums of lognormals, several approximations exist in the literature (Barakat 1976, Beaulieu & Xie 2004, Mehta et al. 2007, Schwartz & Yeh 1982). A numerically stable, very accurate and flexible approach is to approximate the sum of lognormals by a single lognormal random variable (Mehta et al. 2007). Given K independent and lognormally $\mathrm{LN}(\mu, \sigma)$ distributed random variables $U_1, \ldots, U_K$, the Laplace transform of $\sum_{k=1}^{K} U_k$ is matched with the Laplace transform of $U(K)$, the approximating lognormal random variable with parameters $\mu(K)$ and $\sigma(K)$, at two different, real and positive values s_1 and s_2 (Mehta et al. 2007). Thus $\mu(K)$ and $\sigma(K)$ are computed by solving the system of nonlinear equations

$$\sum_{i=1}^{N} \frac{\omega_i}{\sqrt{\pi}} \exp\left[-s_l \exp(\sqrt{2}\sigma(K) a_i + \mu(K)) \right] = [\hat{L}_U(s_l; \mu, \sigma)]^K , \tag{20}$$

for $l = 1, 2$. The choice of s_1 and s_2 is discussed in Section 4. Hence the log-likelihood for the CS model is

$$\sum_{l=1}^{n} \left\{ I_{\{x_l = 0\}}(-\beta t) + I_{\{x_l > 0\}} \log \left[\sum_{m=1}^{\infty} \frac{e^{-\beta(t \wedge t_d)} (\beta(t \wedge t_d))^m}{m!} f_{U(m)}(x_l) \right] \right\} , \tag{21}$$

where $f_{U(m)}$ is the pdf of $U(m)$ with parameters $\mu(m)$ and $\sigma(m)$ computed based on (20).

4. Computational aspects

4.1. Laplace transforms

All computations are carried out using Matlab® (Version 7.4.0). Theoretically, the approximation (13) becomes more accurate the greater N. In practice, there is an optimal value, beyond which numerical error starts to increase the total error (as a function of N) even if the theoretical error of the approximation continuous to

decrease. Set $\hat{L}_U(s) = \hat{L}_U^N(s)$ to make explicit the dependence of the approximation (13) on N. An appropriate N can for example be chosen such that for different parameter settings and different s, the relative error of $\hat{L}_U^N(s)$ in approximating $L_U(s)$ is smaller than 10^{-3} and that the difference $|\hat{L}_U^N(s) - \hat{L}_U^{N+1}(s)|$ is smaller than 10^{-6}. The value of $L_U(s) = \mathbb{E}(\exp(-sU))$ is approximated by simulation. For our implementation, we obtained $N = 20$.

To compute $\hat{L}_{X_t}$ for the shot noise process with $h(t) = \exp(-\lambda t)$, one can now either solve (14) by numerical integration or (15) by using some approximation for $E_1(a)$. Both approaches were carried out by standard routines implemented in Matlab for different parameter settings, and the latter approach was found to be significantly faster with almost identical accuracy to the numerical integration approach.

4.2. Sum of lognormals

There is a trade-off in the choice for s_l $(l = 1, 2)$ when approximating the sum of lognormals (20), in that increasing s_l yields better estimation of the density function for small arguments, whereas reducing s_l yields a better estimation in the tails. As suggested in Mehta et al. (2007), we set $s_1 = 1$ and $s_2 = 0.2$ in (20). This setting provides a reasonable fit for the parameter configurations and values of K that we tested. Figure 1 displays the application of the approach for different values of K. The true distribution of $\sum_{k=1}^{K} U_k$ is approximated by simulations. The estimates are in line with the approximating true distribution. For larger K's, the curve shifts to the right and the total variance increases.

4.3. Likelihood inference

As for N in $\hat{L}_U(s) = \hat{L}_U^N(s)$, we need to find an optimal value for M in $\hat{f}_{X_t,M}$ defined in (17). An appropriate M is chosen by applying the algorithm to test transforms and their (analytical known) inverses given in Table 1 of Abate & Valko (2004), where the test functions in the table have all singularities on the real axis to the left of $s = a$, $a \geq 0$, and are infinitely differentiable. The Laplace transform of one of the test functions is $\exp(-2\sqrt{s})$ which has a comparable curve to our function $\hat{L}_{X_t}(s)$. We obtain $M = 9$ for the present application. Integrating now $\hat{f}_{X_t,M}(x)$ over $(0, \infty)$ and adding the point mass $F_{X_t}(0)$ for different parameter settings yields values which are approximately 1. The results indicate that the approximation implemented works well for our case.

In addition, the performance of the maximum likelihood method for the shot noise process, first with exponential decay (without absorption at zero) and then with constant survival time, is evaluated by estimation from simulated populations with known parameter values of the corresponding model. The size for the simulated dog populations is set to 400 and the ages are drawn from an exponential distribution with mean 3, so that we produce a population somewhat similar to the samples from Kazakhstan, Tunisia and China. Then, based on the specified shot noise process,

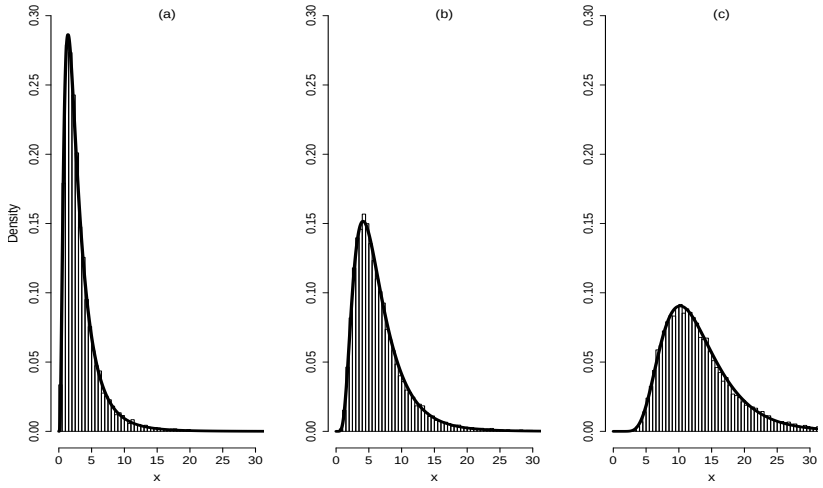

Figure 1: *Approximation (solid line) and the true distribution (histogram based on* 20000 *realizations) for* $\sum_{k=1}^{K} U_k$ *with* U_k*'s independent and lognormally distributed with parameter values* $\mu = 0$ *and* $\sigma = 1$*, with* (a) $K = 2$*,* (b) $K = 4$ *and* (c) $K = 8$*. The parameter* N *in (13) is set to* 20*.*

realizations of the process for the given ages of the dogs are taken as loads for the dogs in the sample. The shot noise process is then fitted by using maximum likelihood to the simulated data. The procedure is repeated 1000 times. Applying this approach with different parameter settings reveals that the methods work well. Representatively, Figure 2 shows the application of the approach to the exponential decay model with fixed parameters $(\beta, \lambda, \mu, \sigma) = (2, 4, 6, 2)$ and for the constant survival model with fixed parameters $(\beta, t_d, \mu, \sigma) = (2, 1, 6, 2)$.

To determine the parameter y_0 in the likelihood function (19) of the PT model, one can successively implement higher derivatives of $\hat{L}_{X_t}$ and test if $|\mathbb{P}(Y_t = y) - \hat{f}_{X_t,M}(y)|$ is smaller than some predefined threshold, say 10^{-4}. Our implementation yields $y_0 = 5$ and thus analytical derivatives up to order 4 are implemented.

5. Application

5.1. Comparison

The MA, PT and CS models are fitted by the maximum likelihood method to the three data sets using the log-likelihoods (18), (19) and (21) respectively. Table 1 shows the resulting estimates. The mean survival times $\mathbb{E}(T)$ in Table 1 are computed based on Subsection 2.5. All three models attest a significantly higher infection pressure β in Kazakhstan and Tunisia than in China, which is reasonable given that a lower prevalence of infection is observed in China (see Figure 3). All

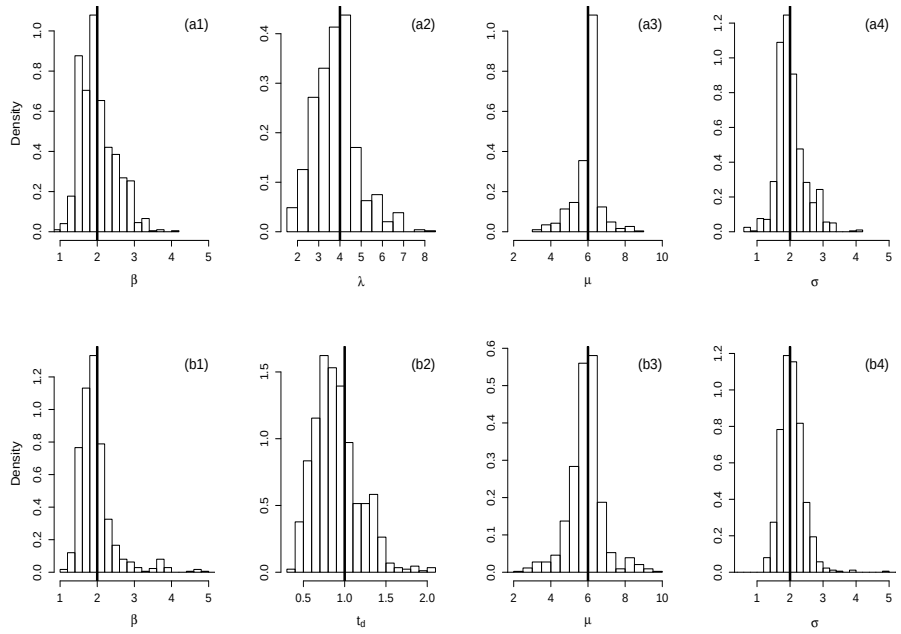

Figure 2: *Histograms of the estimated model parameters from simulated data sets, where* (a1-a4) *the shot noise process with exponential decay with parameters* $(\beta, \lambda, \mu, \sigma) = (2, 4, 6, 2)$ *and* (b1-b4) *the constant survival model with parameters* $(\beta, t_d, \mu, \sigma) = (2, 1, 6, 2)$ *are used to generate the loads for given ages of the dogs for the generation of the samples. The true values are indicated by bold vertical lines.*

models suggest that μ is significantly higher in the Kazakhstan sample than in the others but that σ is similar throughout the samples. The exponential decay rate λ for the MA and PT models is not significantly different in any of the three samples. The mean survival times $\mathbb{E}(T)$, computed based on Subsection 2.5, are lowest for the MA model, but largest for the CS model. They are discussed in more details in the following subsection.

Table 2 shows the observed prevalences of infection q and the means m of the log-transformed observed positive loads of the three samples together with the corresponding model values $\hat{q}$ and $\hat{m}$ computed by simulation as follows. Let n be the sample size and let $t_1, \ldots, t_n$ be the observed ages of dogs in the sample. For $1 \leq k \leq n$, generate a realization for the kth dog with age t_k from the MA, PT or CS model respectively with the parameters set as their estimated values in Table 1, to attribute a simulated load to him. This yields a new sample from which the prevalence and the mean of the log-transformed positive loads can be computed. Repeating the procedure 2000 times, $\hat{q}$ and $\hat{m}$ are computed as the averages of the resulting corresponding 2000 values, and the corresponding 2.5% and 97.5% quan-

Table 1: Maximum likelihood estimates of the mass accumulation model (MA) (equation (6)), the Poisson transform model (PT) (8) and from the constant survival model (CS) (9) for Kazakhstan, Tunisia and China, together with 95% confidence intervals computed by the bootstrap percentile method. The mean survival times $\mathbb{E}(T)$ of a single infection is computed as described in Subsection 2.5. Note that "$-$" indicates that the corresponding parameter is not specified in the model.

	Sample	MA	PT	CS
$\hat{\beta}$	Kaza.	0.501 (0.359, 1.112)	0.445 (0.317, 0.918)	0.340 (0.213, 0.881)
	Tuni.	0.689 (0.417, 1.243)	0.662 (0.401, 1.192)	0.487 (0.312, 0.986)
	China	0.330 (0.226, 0.469)	0.308 (0.197, 0.423)	0.127 (0.014, 0.410)
$\hat{\mu}$	Kaza.	5.474 (4.505, 6.754)	6.001 (4.305, 7.054)	4.302 (3.723, 4.928)
	Tuni.	4.340 (3.096, 5.321)	4.046 (3.462, 5.109)	3.560 (3.101, 4.137)
	China	3.263 (2.535, 4.672)	3.398 (2.403, 4.766)	3.261 (2.718, 3.764)
$\hat{\sigma}$	Kaza.	2.804 (2.417, 3.321)	2.955 (2.437, 3.306)	2.616 (2.182, 2.882)
	Tuni.	2.879 (2.509, 3.471)	3.079 (2.393, 3.399)	2.693 (2.184, 3.157)
	China	2.598 (2.396, 2.988)	2.635 (2.285, 3.108)	2.535 (2.207, 2.890)
$\hat{\lambda}$	Kaza.	9.620 (7.238, 15.617)	8.833 (6.319, 13.176)	$-$
	Tuni.	9.293 (6.783, 18.431)	8.728 (6.210, 16.527)	$-$
	China	8.413 (7.084, 14.981)	7.916 (6.811, 13.672)	$-$
$\hat{t}_d$	Kaza.	$-$	$-$	0.744 (0.580, 1.108)
	Tuni.	$-$	$-$	0.640 (0.397, 1.064)
	China	$-$	$-$	0.713 (0.474, 1.216)
$\mathbb{E}(T)$	Kaza.	0.572	0.745	0.744
	Tuni.	0.476	0.530	0.640
	China	0.403	0.502	0.713

tiles can easily be determined. Overall, the observed values q and m of the models agree with the simulated values of the models. However, for the MA model, the observed values q for the China sample and m for the Kazakhstan sample lie outside the 95% interval of the corresponding simulated quantities.

Figure 3 displays the prevalences of infection and means of the log-transformed positive parasite loads computed by simulation from the MA, PT and CS models for different age classes in all three samples, together with the observed quantities (grey points). The simulation is done as described above. Let $b_{0.025}$ and $b_{0.975}$ denote the 2.5% and 97.5% quantiles from the corresponding simulated values. Given the (simulated) mean prevalence $\hat{q}(t)$ for an age class, the intervals $[b_{0.025}, \hat{q}(t)]$ and $(\hat{q}(t), b_{0.975}]$ were each similar for the three models and in all age classes. Thus instead of plotting all 95% simulation intervals for the three models, we plot the results the averaged interval lengths over the three models at the observed sample values (grey points). Averaging is done over the sub-interval lengths $\hat{q}(t) - b_{0.025}$

Table 2: Observed prevalences of infection q and means of the log-transformed loads m in dogs, together with the corresponding model mean values $\hat{q}$, $\hat{m}$, and 2.5% and 97.5% quantiles, computed by simulation from the MA, PT and CS model as described in the text.

Model	Sample	q	$\hat{q}$	m	$\hat{m}$
MA	Kaza.	0.230	0.243 (0.208, 0.278)	4.503	3.682 (3.331, 4.061)
	Tuni.	0.271	0.277 (0.193, 0.357)	3.826	3.283 (2.611, 4.005)
	China	0.084	0.125 (0.096, 0.156)	3.322	2.594 (1.891, 3.409)
PT	Kaza.	"	0.237 (0.204, 0.273)	"	4.265 (3.843, 4.699)
	Tuni.	"	0.261 (0.191, 0.331)	"	3.548 (2.827, 4.291)
	China	"	0.116 (0.081, 0.148)	"	2.939 (2.206, 3.738)
CS	Kaza.	"	0.211 (0.173, 0.259)	"	4.685 (4.301, 5.077)
	Tuni.	"	0.245 (0.174, 0.309)	"	4.259 (3.505, 5.082)
	China	"	0.078 (0.051, 0.105)	"	3.802 (3.022, 4.691)

and $b_{0.975} - \hat{q}(t)$, and the resulting (averaged) values are then plotted around the observed quantity (grey bars). This gives a reasonable indication of the variation of the models. Analogously, we proceed for the case of the mean load $\hat{m}(t)$ of the log-transformed positive loads in all age classes. The results in Figure 3 indicate that all models suggest an asymptotic prevalence of infection much lower than 1 in the three samples, and a decreasing mean load of the log-transformed positive loads in young dogs which stabilizes after about 1 year. The PT and CS models are well in line with the observed quantities. In plot (b1), the mean prevalence of the MA model for the age class 6.5+ is slightly larger than the upper bound of the (averaged) simulation interval. In plot (a2), the mean load of the log-transformed positive loads computed by the MA model is close to the lower bounds of the simulation intervals in age classes $(0.5, 1.5]$, $(1.5, 2.5]$ and $(3.5, 4.5]$, and even slightly below in age class $(2.5, 3.5]$. The MA model produces the highest prevalences and the lowest means of the log-transformed positive loads in all samples.

5.2. Bounding models

The results in Table 2 and Figure 3 suggest that the PT and CS perform better than the MA model. Figure 4 shows a plot of the estimated pdfs of the PT and CS models for the log-transformed positive loads in the Kazakhstan sample for different age classes, together with a histogram of the corresponding observed quantities. The Kazakhstan sample is chosen since it has the largest number of observations of positive loads. For simplicity, the densities are computed by simulation. Given the middle point of an age class t, we generate 100000 realizations of the PT respectively CS model for t fixed, and take the logarithms of the positive loads. Then a kernel-estimator is applied to the log-transformed loads to obtain an approximation to the

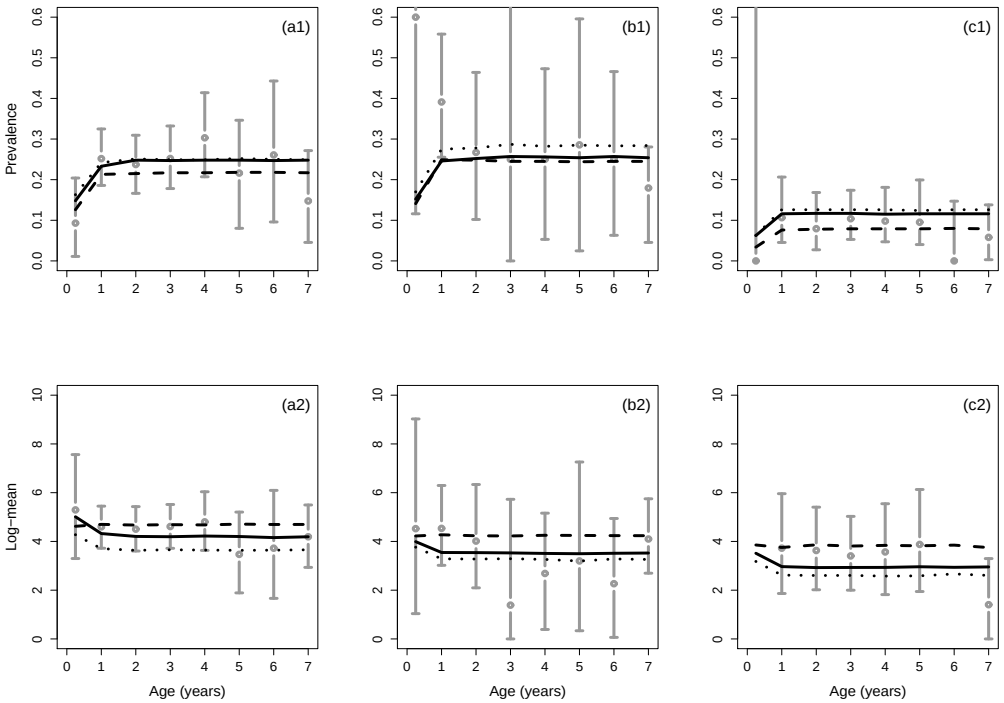

Figure 3: *Prevalences of infection* (a1-c1) *and means of the log-transformed positive loads* (a2-c2) *in dogs from the MA model (dotted line), PT model (solid line) and CS model (dashed line) (computed by simulation) versus the observed quantities (grey points) for the age classes* $(0, 0.5]$, $(0.5, 1.5]$, $(1.5, 2.5]$, $(2.5, 3.5]$, $(3.5, 4.5]$, $(4.5, 5.5]$, $(5.5, 6.5]$ *and* $6.5+$ *of the samples* (a)*Kazakhstan* , (b)*Tunisia and* (c)*China. The grey bars indicate the averaged* 95% *simulation intervals for the three models, computed as described in the text. Note that there are no observed positive loads in age classes* $(0, 0.5]$ *and* $(5.5, 6.5]$ *of the China sample and hence the simulation intervals are not plotted here.*

true pdf, resulting in the solid and dashed lines in Figure 4. The fitted models provide a reasonable description of the positive parasite loads.

Since it is likely that the true decay dynamics for ingested parasite clumps in dogs is intermediate between the PT and CS models, the values of quantities such as infection pressure, mean load and mean survival time of a clumped infection obtained from these models should indicate a plausible range of values for these quantities. Hence averaging the corresponding values of the PT and CS models is a reasonable choice to summarize the results. Using Table 1, the averaged infection rates β are 0.393, 0.575 and 0.218 infections per dog per year in Kazakhstan, Tunisia and China respectively. This could indicate that the prevalences of infection of sheep

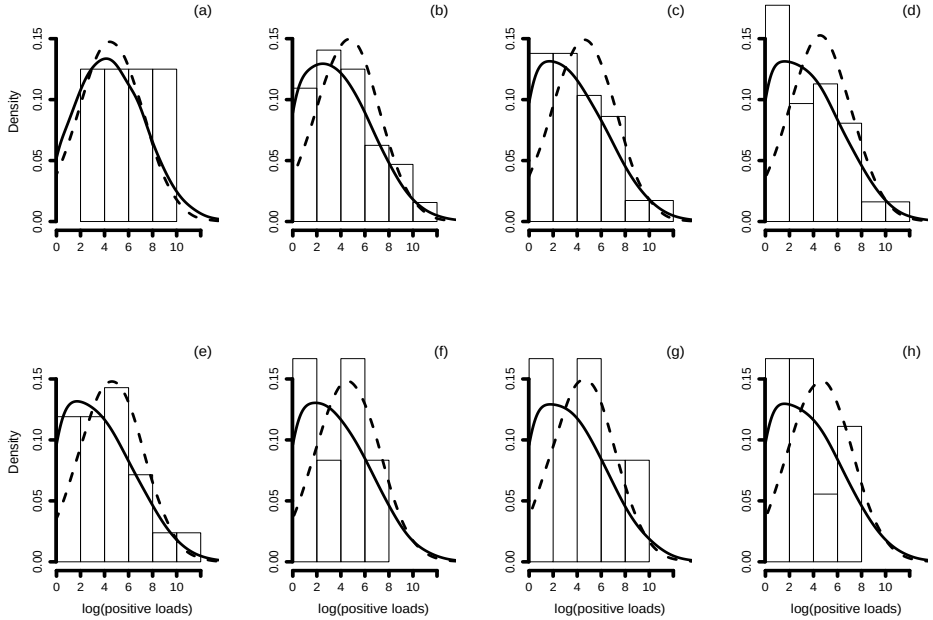

Figure 4: *Plot of the estimated density functions for the log-transformed positive loads in the Kazakhstan sample by the PT model (solid line) and the CS model (dashed line) for the age classes* (a) $(0, 0.5]$, (b) $(0.5, 1.5]$, (c) $(1.5, 2.5]$, (d) $(2.5, 3.5]$, (e) $(3.5, 4.5]$, (f) $(4.5, 5.5]$, (g) $(5.5, 6.5]$ *and* (h) 6.5+, *together with a histogram of the observed log-transformed positive loads for the above age classes. Note that there are* 4, 32, 29, 31, 21, 8, 6 *and* 9 *observed positive loads in the above age classes.*

in Kazakhstan and China are lower than in Tunisia, or that dogs in Kazakhstan and China consume less viscera of sheep.

The average of the values of the mean μ of the log-transformed positive loads is with 5.152 in Kazakhstan larger than in Tunisia with 3.803 and China with 3.330. The averaged values for σ of 2.806, 2.886 and 2.585 for the Kazakhstan, Tunisia and China samples respectively are comparable. The corresponding mean clump sizes of a single infection are then 9000, 3000 and 1000 parasites for the Kazakhstan, Tunisia and China sample. The higher average clump size in Kazakhstan could be due to a higher number of fertile cysts in infected sheep, or that the age of sheep at slaughter are higher, or that there is a higher infection pressure in sheep from Kazakhstan, implying that cysts are acquired by sheep at younger ages and thus have more time to develop and become fertile.

The averaged mean durations of a single infection are with 0.745, 0.585 and 0.608 years for the Kazakhstan, Tunisia and China samples respectively similar. Hence we have an overall mean survival time of about 0.65 years $\approx$ 8 months, which is in

line with the results in Gemmell (1959), who suggested that the mean duration of infection is slightly lower than 1 year. Our estimate is also in line with the about 10 months suggested in Aminzhanov (1975). They experimentally infected 48 $3-4$ month old dogs with 25000 protoscoleces, and killed groups of 4 dogs at times regularly distributed over the year. However, our estimate is lower than the 18 months derived in Roberts et al. (1986) from a prevalence model fitted to data from New South Wales.

6. Discussion

In this paper, different shot noise processes are used to describe the ingestions of clumps containing *Echinococcus granulosus* parasites in dogs. The processes model clumped superinfections coupled with a time-dependent decay mechanism of the ingested parasite burden, namely an exponential decay with absorption around zero (MA model), a Poisson transform of the exponential shot noise process (PT model) and a constant duration of infection (CS model). Based on the skewness in the data, a lognormal distribution is chosen as distribution for the number of parasites per clump.

Maximum likelihood estimation is used to fit the models to samples from Kazakhstan, Tunisia and China. The PT and CS model are shown to perform best. Since the true decay dynamics for ingested parasite clumps in dogs most likely lies between the PT and CS model dynamics, their parameter estimates are averaged.

The results suggest that the infection rate is about 0.4, 0.6 and 0.2 infections per dog per year in Kazakhstan, Tunisia and China respectively. The lower values in Kazakhstan and China could be the result of a lower consumption of infected viscera of sheep, due to a lower prevalence of infection in sheep or a different feeding behavior of dogs as compared to Tunisia. The mean number of parasites in a clumped infection is about 9000 in Kazakhstan, 3000 in Tunisia and and 1000 China. The higher average clump size in Kazakhstan could be due to a higher number of fertile cysts in infected sheep, or that the age of sheep at slaughter are higher, or that there is a higher infection pressure in sheep from Kazakhstan, implying that cysts are acquired by sheep at younger ages and thus have more time to develop and become fertile. Hence the infection of dogs with *Echinococcus granulosus* occur at a low rate, but the ingested parasite load per clump is in the thousands. The mean duration of a single clumped infection is about 8 months, comparable in all three samples. The value is in line with other studies, Gemmell (1959) suggesting a mean time of slightly lower than 1 year, and Aminzhanov (1975) suggesting a mean time of about 10 months.

Achnowledgements The work was supported by the Schweizerischer Nationalfonds (SNF), project no. 107726.

References

Abate, J. & Valko, P. P. (2004), 'Multi-precision Laplace inversion', *Int. J. Numer. Meth. Engng.* **60**, 979–993.

Abramowitz, M. & Stegun, I. A. (1972), *Handbook of mathematical functions with formulas, graphs and mathematical tables*, 9 edn, Dover.

Aminzhanov, M. (1975), 'Duration of the life of Echinococcus granulosus in the organism of dogs', *Veterinariia* **12**, 70–72.

Anderson, R. M. & May, R. M. (1978), 'Regulation and stability of host parasite population interactions, i, Regulatory processes', *J Anim Ecol* **47**, 219–249.

Balling, T. E. & Pfeiffer, W. (1997), 'Frequency distributions of fish parasites in the perch *Perca fluviatilis* l. from Lake Constancee', *Parasitol Res* **83**, 370–373.

Barakat, R. (1976), 'Sums of independent lognormally distributed random variables', *J. Opt. Soc. Am.* **66**, 211–216.

Beaulieu, N. C. & Xie, Q. (2004), 'An optimal lognormal approximation to lognormal sum distributions', *IRE Trans. Commun. Syst.* **53**, 479–489.

Braga, C., Ximenes, R., Miranda, J. & Alexander, N. (2005), 'Bancroftian filariasis in an endemic area of Brazil: differences between genders during puberty', *Rev. Soc. Bras. Med. Trop.* **38**, 224–228.

Budke, C. M., Qiu, J., Craig, P. S. & Torgerson, P. R. (2005), 'Modeling the transmission of Echinococcus granulosus and Echinococcus multilocularis in dogs for a high endemic region of the Tibetan plateau', *Int J Parasitol.* **35**, 163–170.

Cox, D. R. & Isham, V. (1980), *Point Processes*, 2 edn, New York: Chapman and Hall.

Das, P. K., Subramanian, S., Manoharan, A., Ramaiah, K. D., Vanamail, P., Grenfell, B. T., Bundy, D. A. P. & Michael, E. (1995), 'Frequency distribution of Wuchereria bancrofti infection in the vector host in relation to human host: evidence for density dependence', *Acta Tropica* **60**, 159–165.

Eckert, J. & Deplazes, P. (2004), 'Biological, epidemiological and clinical aspects of Echinococcosis, a zoonosis of increasing concern', *Clin Microbiol Rev.* **17**, 107–135.

Economides, P. & Cristofi, G. (2002), *Cestode zoonoses: Echinococcosis and cysticercosis. An emergent and global problem*, 3 edn, NATO Science Series:IOS Press Amsterdam.

Gemmell, M. A. (1959), 'Hydatid diseases in Australia. IV. Observations on the incidence of Echinococcus granulosus on stations and farms in endemic regions of New South Wales', *Aust Vet J* **35**, 396–402.

Gemmell, M. A., Lawson, J. R. & Roberts, M. G. (1986), 'Population dynamics in echinococcosis and cysticercosis: biological parameters of Echinococcus granulosus in dogs and sheep', *Parasitology* **92**, 599–620.

Grenfell, B. T., Wilson, K., Isham, V. S., Boyd, H. E. G. & Dietz, K. (1995), 'Modelling patterns of parasite aggregation in natural populations: trichostrongylid nematode-ruminant interactions as a case study', *Parasitology* **111(Suppl.)**, S.135–151.

Heinzmann, D., Barbour, A. D. & Torgerson, P. R. (2009), 'Compound processes as models for clumped parasite data', *J. Appl. Probab.* **222**, 27–35.

Herbert, J. & Isham, V. (2000), 'Stochastic host-parasite interaction models', *J. Math. Biol.* **40**, 343–371.

Irvine, R. J., Stien, A., Dallas, J. F., Halvorsen, O., Langvatn, R. & Albon, S. D. (2000), 'Life-history strategies and population dynamics of abomasal nematodes in Svalbard reindeer (Rangifer tarandus platyrhynchus)', *Parasitol* **120**, 297–311.

Kapel, C. M. O., Torgerson, P. R., Thompson, R. C. A. & Deplazes, D. (2006), 'Reproductive potential of Echinococcus multilocularis in experimentally infected foxes, dogs, raccoon dogs and cats', *Int J Parasitol.* **36**, 79–86.

Lahmar, S., Kilani, M. & Torgerson, P. R. (2001), 'Frequency distributions of Echinococcus granulosus and other helminths in stray dogs in Tunisia', *Ann Trop Med Parasitol* **95**, 69–76.

Lund, R. B., Butler, R. W. & Paige, R. L. (1999), 'Prediction of shot noise', *J. Appl. Prob.* **36**, 374–388.

Mehta, N. B., Wu, J., Molisch, A. F. & Zhang, J. (2007), 'Approximating a sum of random variables with a lognormal', *IEEE Trans Wireless Comm* **6**, 2690–2699.

Nodtvedt, A., Dohoo, I., Sanchez, J., Conboy, G., DesCôteaux, L., Keefe, G., K., L. & Campell, J. (2002), 'The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows', *Can J Vet Res.* **66**, 249–257.

Pugliese, A., Rosa, R. & Damaggio, M. L. (1998), 'Analysis of a model for macroparasitic infection with variable aggregation and clumped infections', *J Math Biol* **36**, 419–447.

Roberts, M. G., Lawson, J. R. & Gemmell, M. A. (1986), 'Population dynamics in echinococcosis and cysticercosis: Mathematical model of the life-cycle of Echninococcus granulosus', *Parasitology* **92**, 621–641.

Schwartz, S. & Yeh, Y. (1982), 'On the distribution function and moments of power sums with lognormal components', *IRE Trans. Commun. Syst.* **8**, 1441–1462.

Stehfest, H. (1970), 'Algorithm 368: numerical inversion of laplace transforms', *Commun. ACM* **13**, 47–49.

Stoer, J. & Bulirsch, R. (1983), *Introduction to numerical analysis*, 2 edn, Springer: Berlin.

Thompson, R. C. A. & Lymbery, A. J. (1986), *The biology of Echinococcus and hydatid disease*, London: George Allen and Unwin.

Torgerson, P. R., Burtisurnov, K. K., Shaikenov, B. S., Rysmukhambetova, A. T., Abdybekova, A. M. & Ussenbayev, A. E. (2003a), 'Modelling the transmission dynamics of Echinococcus granulosus in dogs in rural Kazakhstan', *Parasitology* **126**, 417–424.

Torgerson, P. R., Oguljahan, B., Muminov, M. E., Karaeva, R. R., Kuttubaev, O. T., Aminjanov, M. & Shaikenov, B. (2006), 'Present situation of cystic echinococcosis in Central Asia', *Parasitol Int.* **55**, 207–212.

Torgerson, P. R., Ziadinov, I., Aknazarov, D., Nurgaziev, R. & Deplazes, P. (2009), 'Modelling the age variation of larval protoscoleces of echinococcus granulosus in sheep', *Int J Parasitol.* **39**, 1031–1035.

Widder, D. V. (1946), *The Laplace transform*, Princton: Princton University Press.

Woolhouse, M. E. J., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L. K., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K. & Anderson, R. M. (1997), 'Heterogeneities in the transmission of infectious agents: Implications for the design of control programs', *PNAS* **94**, 338–342.

A mechanistic two-host model for the transmission of *Echinococcus granulosus*

Dominik Heinzmann[1,2], A.D. Barbour[1], and Paul R. Torgerson[2,3]
[1]Institute of Mathematics, University of Zurich, Switzerland
[2]Institute of Parasitology, University of Zurich, Switzerland
[3]School of Veterinary Medicine, Ross University, West Indies

Abstract

A stochastic process for the infection dynamics of the parasite *Echinococcus granulosus* in a two-host transmission system is proposed. The model describes the densities of the parasite in the host populations. The architecture consists of two sub-models for the acquisition and severity of infection in the host populations and a superposed infection contact scheme. The parasite dynamics within the host population are modeled using a compound mixed Poisson process for the sheep and a shot-noise process for the dogs. A threshold of extinction is derived. The model output in the stationary setting is shown to reflect reasonably the observed parasite distributions in the hosts. The sensitivity of the model to environmental factors and control interventions is investigated.

Keywords: Shot noise process, clumped infection, basic reproduction number, generation time, *Echinococcus*.

1. Introduction

A stochastic process is used to model a parasitic two-host transmission system. The model describes the development of the *Echinococcus granulosus* (Eckert & Deplazes 2004) parasite densities in the definitive host, a dog, and the intermediate host, a sheep. The parasite causes cystic echinococcosis which is a zoonotic parasitic disease, endemic in many parts of the world (Economides & Cristofi 2002, Torgerson et al. 2006). Adult worms mature in the intestine of the dog and the infective eggs are released in the feces. Conditional on ingestion of such infective biomass by sheep, hydatid cysts can form in organs such as the liver ($60 - 70\%$), lungs and brain. The cyst develops over years in the sheep. Dogs gain access to infective materials by consumption of infected viscera of sheep. In rural areas, sheep are slaughtered on the farms and the offal is accessible to dogs. Dogs can also become infected by scavenging dead sheep in fields.

Roberts et al. (1986) constructed a model of the *Echinococcus granulosus* life-cycle. They used integrodifferential equations for the mean number of worms in dogs

and cysts in sheep. The host populations were divided into classes according to the number of infections they carry in order to focus on acquired immunity in hosts. Animals were assumed to lose infection independently of the parasite burden. The model mechanistically described the prevalence and the development of the mean burden of parasites in the host, but not the parasite densities. A negative binomial distribution was used to describe the densities in hosts, but without linking it to an underlying infection process or to the animals' ages.

In this paper, a mechanistic model describing the development of the worm loads in dogs and cyst burdens in sheep is presented. The architecture is based on the two sub-processes introduced in Heinzmann et al. (2009, n.d.) to model the infection dynamics in the two host populations separately. In Heinzmann et al. (2009), a compound mixed Poisson process is used to model the acquisition of hydatid cysts in sheep, whereas in Heinzmann et al. (n.d.), a shot noise process is used to model the infection pattern in dogs. The model here links these two models by superposing a biologically reasonable infection contact pattern between the hosts, yielding a model for the whole life-cycle of the parasite. All parameters of the model are estimated based on observed data. The model describes the age-dependent prevalence and distribution of worms in dogs and cysts in sheep, and includes the age-dependent fertility of cysts as an element. Simulation of the full model shows that key quantities such as the prevalences of infection and the infection pressures are well in line with observed data. A basic reproduction number is derived as a function of the parameters, indicating that, in the area from which our data originated, a single infected dog in fully susceptible dog and sheep populations (indirectly) causes on average 1.8 new infections in the dog population.

The mechanistic model is then used to evaluate the influence of environmental factors and control intervention programs on the transmission cycle. The simulation experiments show that the model provides a valuable tool for investigating the dynamics initiated by natural or man-made changes to the life-cycle of *Echinococcus granulosus*.

2. Clumped infection processes as sub-models

In this section, the sub-models for the sheep and dog populations proposed in Heinzmann et al. (2009, n.d.) are briefly recalled. The parameters of those models are fixed based on the estimates obtained from *Echinococcus granulosus* samples from Kazakhstan, since for this country, both dog and sheep samples are available. The sheep sample (Torgerson et al. 2003b) contains 2505 individual records of the variables age and hydatid cyst burden in sheep. The dog sample contains 606 individual records of the variables age and parasite counts in dogs (Torgerson et al. 2003a).

The sub-models are based on a clumped infection mechanism. It is assumed in both models that the transmission cycle of *Echinococcus granulosus* is endemically stable, so that the ingested parasite clumps are identically distributed over time.

Endemicity of the transmission cycle is suggested by various studies (Gemmell et al. 1986, Torgerson et al. 1998, Todorov & Boeva 1999, Torgerson et al. 2003a,b, Ziadinova et al. 2008). Furthermore, a low incidence rate in hosts is also assumed in both models so that the acquisition process can be described by a (mixed) Poisson process. A low infection rate of sheep is suggested by Cabrera et al. (1995), Gemmell et al. (1986) and Torgerson et al. (2003b), and a low infection rate of dogs is suggested by Gemmell (1959), Roberts et al. (1986), Torgerson et al. (2003a) and Torgerson et al. (2006). Finally, the clump sizes are treated as independent in the models, making compound processes for sheep and shot noise processes for dogs reasonable models.

2.1. Infection dynamics in sheep population

In Heinzmann et al. (2009), it is shown that a compound mixed Poisson process with a zero-truncated negative binomial distribution for the number of established cysts per ingested clump of infective material provides an adequate fit for the age-dependent cyst distribution in sheep. The model indicates that the rate of ingestion of clumps is heterogeneous within the sheep population, and that the clump sizes are aggregated.

Let the random variable Y_t be the total number of cysts established in an individual up to time t and let V_k ($k = 1, 2, \ldots$) be independent random variables describing the numbers of cysts acquired at a single infection, with common distribution $\mathcal{Q}$. $\mathcal{Q}$ is assumed to be the zero-truncated version of a negative binomial distribution with parameters θ and ζ, whose mean and variance are given by

$$\mathbb{E}(V_k) = \frac{\theta\zeta}{1-(1/(\zeta+1))^\theta}$$

and

$$\mathrm{Var}(V_k) = \left[\frac{\theta\zeta(1+\zeta+\theta\zeta)}{1-(1/(\zeta+1))^\theta} - \left(\frac{\theta\zeta}{1-(1/(\zeta+1))^\theta}\right)^2\right].$$

Denote the distribution of Y_t by $\mathcal{P}_t$. Then the model presumes that

$$Y_t = \sum_{k=1}^{N_t} V_k \quad \text{with} \quad \mathbb{P}(N_t = m) = \frac{\Gamma(\psi+m)}{\Gamma(\psi)m!}\left(\frac{1}{t\xi+1}\right)^\psi\left(\frac{t\xi}{t\xi+1}\right)^m,$$

so that

$$\mathcal{P}_t = \sum_{m=0}^{\infty} \mathbb{P}(N_t = m)\mathcal{Q}^{*m}, \tag{1}$$

where $\mathcal{Q}^{*m}$ is the mth convolution of $\mathcal{Q}$ and N_t is independent of the V_k's. N_t is thus a negative binomial random variable with mean $\psi\xi t$ and variance $\psi\xi t(1+\xi t)$, where ψ is a shape and ξ a scale parameter. The model was fitted to a sheep sample from Kazakhstan, leading to the parameter estimates represented in Table 1. The estimates imply that a sheep ingests an infectious clump roughly every $1/\hat{\psi}\hat{\xi} \approx 3$ years, each clump leading on average to about 4 established cysts.

Table 1: Maximum likelihood estimates of the parameters of the sub-process (1) for the sheep population and of the PT respectively CS sub-process for the dog population from the corresponding Kazakhstan samples, together with 95% confidence intervals computed by the bootstrap percentile method. Note that " − " for the dog models indicates that the corresponding parameter is not specified in that model.

Sheep	Sub-process	
$\hat{\psi}$	$0.941\,(0.629, 1.260)$	
$\hat{\xi}$	$0.343\,(0.225, 0.741)$	
$\hat{\theta}$	$0.351\,(0.139, 0.617)$	
$\hat{\zeta}$	$5.859\,(3.215, 9.763)$	
Dogs	PT-subprocess	CS-subprocess
$\hat{\beta}$	$0.445\,(0.317, 0.918)$	$0.340\,(0.213, 0.881)$
$\hat{\mu}$	$6.001\,(4.305, 7.054)$	$4.302\,(3.723, 4.928)$
$\hat{\sigma}$	$2.955\,(2.437, 3.306)$	$2.616\,(2.182, 2.882)$
$\hat{\lambda}$	$8.833\,(6.319, 13.176)$	−
$\hat{t}_d$	−	$0.744\,(0.580, 1.108)$

2.2. Infection dynamics in dog population

Heinzmann et al. (n.d.) proposed shot noise processes as models for the distribution of worms in dogs, with different time-dependent decay dynamics of the established parasite burdens. They were fitted to three different data sets, and the estimates were shown to be plausible when compared to experimental data. Simulation studies showed that the models satisfactorily reflect the prevalence of infection and the mean of the log-transformed loads for the samples.

Let the random variable X_t denote the total number of parasites carried by a dog at time t. $(X_t)_{t\geq 0}$ is modeled as a continuous-time piecewise deterministic stochastic process. Infection events occur at the times $0 < \tau_1 < \tau_2 \ldots$ of a Poisson process with rate β. At each τ_i, the value of X_t increases by a random amount, assumed to have a log-normal distribution $\mathrm{LN}(\mu, \sigma^2)$. Thereafter, the amount declines according to a predetermined scheme. Thus

$$X_t = \sum_{k=1}^{M_t} U_k h(t - \tau_k)\,, \quad t \geq 0\,, \tag{2}$$

where M_t is a Poisson random variable with mean βt, $U_k \sim \mathrm{LN}(\mu, \sigma^2)$ $(k = 1, 2, \ldots)$ are independent, and $h(t)$, $t \geq 0$, denotes the proportion of parasites still surviving t time units after infection. We take $h(t) = 0$ for $t < 0$.

There is experimental evidence that dogs eventually completely lose their infection (Aminzhanov 1975, Eckert & Deplazes 2004, Gemmell et al. 1986); this should

also be reflected in the model. The natural choice $h(t) = \exp(-\lambda t)$, for some $\lambda > 0$, is not of this form. For this version, we prefer a more faithful model, in which the number of parasites $\tilde{U}_k$ acquired by a dog at the kth infection has a mixed Poisson distribution $\mathrm{Po}(U_k)$, where $U_k \sim \mathrm{LN}(\mu, \sigma^2)$ as before, and each of the $\tilde{U}_k$ parasites has an independent, exponentially distributed lifetime with mean $1/\lambda$. Then an infection load of m parasites disappears completely after a random time of mean $(\gamma + \log m)/\lambda$, where γ is the Euler-Mascheroni constant. For $U_k \sim \mathrm{LN}(\mu, \sigma^2)$, this gives a mean survival time of a single infection load of about $(\gamma + \mu)/\lambda$. Because of the mixed Poisson assumption, the number of parasites X_t also has a mixed Poisson distribution $\mathrm{Po}(\sum_{k=1}^{M_t} U_k \exp(-\lambda(t - \tau_k)))$. This model is referred to as the PT (Poisson transform) model.

An alternative is offered by the CS (constant survival) model, where

$$h(t) = \begin{cases} 1 & \text{if } t \le t_d \\ 0 & \text{if } t > t_d \,, \end{cases}$$

with t_d the (fixed) duration of infection in a dog. Here, there is no decay of the burden between infection and complete loss at t_d time units later. The PT and CS models can be thought of as extreme cases, where the true decay dynamics lying somewhere in between. Hence we average the corresponding estimates given in Table (1) obtained by fitted the PT and CS models to a dog sample of size 606 from Kazakhstan. The results suggest that the infection rate is about 0.4 per dog per year, indicating that a dog is exposed to infection on average once every 2.5 years, with a mean load of $\exp(\mu + \sigma^2/2) = \exp(5.152 + 2.786^2/2) \approx 8500$ parasites per infection. The mean survival time of a single infection in dogs of the PT model is $(\gamma + 6.001)/8.833 = 0.745$ years. Thus the two models suggest that a single clumped infection survives for about 9 months.

3. Fertility of cysts

Heinzmann et al. (2009) considered the age-dependent distribution of cysts in sheep. Since only cysts containing protoscoleces are fertile and can thus lead to an infection in dogs, the distribution of protoscoleces in cysts needs also to be investigated.

3.1. Data set

The distribution of the protoscoleces in fertile cysts is derived from unpublished data from Kyrgyzstan, collected in 2007. A total of 1081 sheep slaughtered in an abattoir were examined for *Echinococcus granulosus* cysts. Conditional on presence of cysts, the number of protoscoleces was evaluated by individual counting of the fertile cysts (i.e. cysts containing protoscoleces). The radii were (arbitrary) recorded for approximately 30% of the fertile cysts. For around 90% of those sheep harboring multiple cysts with at least two of them fertile, only the pooled counts of protoscoleces

Table 2: Number of cysts n_1, number of fertile cysts n_2 and the resulting observed proportion of fertile cysts $q(t)$ (with 95% binomial confidence interval) for different ages. Ages were recorded as integers in the sample with 6 years the maximum.

Age	n_1	n_2	$q(t)$
1	529	5	0.009 (0.003, 0.022)
2	810	14	0.017 (0.009, 0.029)
3	662	34	0.051 (0.036, 0.071)
4	783	44	0.056 (0.041, 0.075)
5	617	38	0.062 (0.044, 0.084)
6	490	43	0.088 (0.064, 0.116)
overall	3891	178	0.046 (0.039, 0.053)

were available. Such counts make up 39.4% of the total protoscolex number in the sample and hence it is necessary to include them in the analysis. This can be done by using their radii, which were recorded for all fertile cysts in the data having multiple cyst counts. The cysts for which both radius r and protoscolex count were recorded make up 20% of the sample, and a linear regression of protoscolex count against r^3 yields a slope of $\hat{a} = 528.480$, with reasonable goodness-of-fit verified by different diagnostic tools such as residual analysis. Hence the pooled counts where the radii of the fertile cysts are available can be divided between the fertile cysts, in proportion to their r^3 values.

The final data set used for the analysis contains those sheep having cysts, either not fertile or individually counted or with counts approximated as described above, together with the age of the sheep. From a total of 661 sheep with a total number of 3891 cysts (fertile and non-fertile), 606 sheep do not have any fertile cysts, and the remaining 55 harbor 178 fertile cysts. Table 2 summarizes the observed counts and the resulting proportions of cysts having protoscoleces for different age classes of the sheep. The overall prevalence of infection is 0.046, indicating a low fertility of cysts in general. The histogram of the log-transformed protoscolex burdens of fertile cysts in Figure 2 indicates that most burdens are large.

3.2. Two-part conditional model

The distribution of protoscoleces in cysts has an excess of zeros and the remaining positive counts are positively skewed. It is reasonable to assume that the initial time points when cysts start developing protoscoleces and the protoscolex population growth inside the cysts are governed by different processes, so that a two-part conditional model (Cohen 1960) can be used for such data (Duan et al. 1984).

Let Z be a random variable which models the protoscolex burden in cysts. The

two-part conditional scheme is

$$\begin{aligned} \mathbb{P}(Z=0) &= 1-q \\ \mathbb{P}(Z=z) &= qf(z) \ , \ z=1,2,\dots , \end{aligned} \tag{3}$$

where q is the probability of cyst fertility in a sheep of age t and $f(y)$ is a zero-truncated probability mass function describing the nonzero burdens.

Let $z=(z_1,\dots,z_m)$ be the counts arising from a random sample of m cysts in sheep aged $t_1,\dots,t_m$. The log-likelihood function of model (3) can be written as

$$\begin{aligned} l(z) = &\textstyle\sum_{j=1}^{m}\{I_{\{z_j=0\}}\log(1-q_j)+(1-I_{\{z_j=0\}})\log(q_j)\} \\ &+\textstyle\sum_{j=1}^{m} I_{\{z_j>0\}} f(z_j)\,, \end{aligned} \tag{4}$$

where I is the indicator function, q_j denotes the value of q for cysts in a sheep of age t_j and f is a zero-truncated distribution. This decomposition of the log-likelihood function l shows that q and f can be estimated and interpreted separately (Welsh et al. 1996).

3.3. Age-dependent fertility

Let $k(t)$ be the probability that a cyst of age t has formed protoscoleces and thus is fertile. Assume that once a cyst is fertile, it remains so. Since $k(0)=0$ and the shape of $k(t)$ should allow a flexible fit, a reasonable choice for $k(t)$ is the Bass model (Bass 1969):

$$k(t) = k^* \frac{1-e^{-(a+b)t}}{1+\frac{b}{a}e^{-(a+b)t}}\,, \tag{5}$$

where k^* is the asymptotic probability of fertility and a and b are adjustable coefficients. Equation (5) allows the modeling of S-shaped curves.

Our data contain records of the number of fertile and non-fertile cysts for a sheep of age t, but not the age of the cyst itself. Thus (5) is a latent process and needs to be coupled to the underlying mixed Poisson infection process of (1) in order to compare it to the data. Each animal has a fixed infection rate, and the acquisition process of cysts is Poisson. Let $q(t)$ denote the probability that a cyst is fertile in a sheep of age t. Thus the following equation for $q(t)$ is appropriate to model the fertility in cysts:

$$\begin{aligned} q(t) &= \int_0^t k(s)\frac{1}{t}ds = \frac{k^*}{t}\int_0^t \frac{1-e^{-(a+b)s}}{1+\frac{b}{a}e^{-(a+b)s}}ds \\ &= \frac{k^*}{t}\left[t+\frac{1}{b}\left\{\log(1+\frac{b}{a}e^{-(a+b)t})-\log(1+\frac{b}{a})\right\}\right]. \end{aligned} \tag{6}$$

The maximum likelihood estimates of the parameters in (6) are $\hat{k}^*=0.103$ (95% bootstrap confidence interval: $(0.079, 0.121)$), $\hat{a}=0.124$ $(0.101, 0.147)$ and $\hat{b}=1.394$

(1.204, 2.426). Figure 1 displays the resulting latent model (5) and the mixture model (6). The latent model suggests that a cyst has a chance of about 6% to be fertile at age 2 and a chance of 10% at age 4. Between $1-3$ years, cyst fertility increases approximately linearly with age. The mixture model (6) suggests that the asymptotic proportion of fertile cysts in older sheep is approximately 10%, indicating that one cyst out of ten is on average fertile. Figure 1(b) shows that the mixture model is in reasonable agreement with the observed proportions of cyst fertility in different age classes.

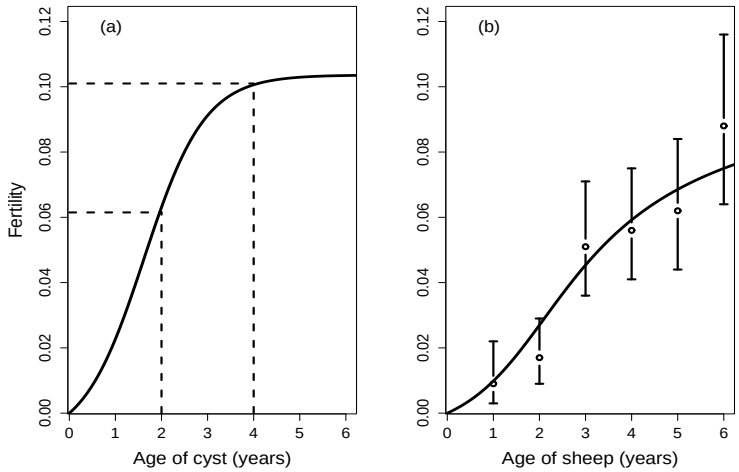

Figure 1: (a) *Estimated age-dependent fertility equation (5) for Echinococcus granulosus cysts.* (b) *The resulting fertility of cysts in function of the age of the sheep based on the mixture model (6), together with the observed prevalences of infection given in Table 2.*

3.4. Distribution of protoscoleces in fertile cysts

For the sub-model (2) of the dog population, we have specified the clump distribution as being lognormal. Since a two-part conditional model is used to describe the distribution of protoscoleces in cysts and since the probability of fertility of cysts is low, it can be assumed that if the dog eats a sheep infected with cysts, it only ingests a small number of fertile cysts (rarely larger than 1). Thus it is reasonable to check if the protoscolex distribution in fertile cysts is lognormal. Figure 2 shows the histogram and a quantile-quantile normal plot from the log-transformed protoscolex burdens in fertile hydatid cysts, indicating that the choice of log-normality in the positive protoscolex burdens is reasonable. The maximum likelihood estimates of the mean and standard deviation are 7.444 ($95\%CI$: 7.049, 7.834) and 2.010 ($95\%CI$: 1.782, 2.313) respectively. The resulting distribution is given in Figure 2(a). The

mean and standard deviation estimates are different from the estimates given in Table 1 of the parameters μ and σ of the lognormal distribution, which is used in the PT and CS models to determine the number of parasites acquired by a dog at infection. The different values may arise because not all protoscoleces are transformed into worms. The number of established worms is likely to depend on the individual immune system of dogs and thus could explain the increase in variance of the ingested clump distribution. Gemmell et al. (1986) have shown that the distribution of established parasites conditional on a single infection with 17500 protoscoleces is heavily skewed, ranging from a small load to several thousands. In addition, dogs can consume more than one fertile cyst per ingestion, which could also increase the variance in the number of established parasites per infection.

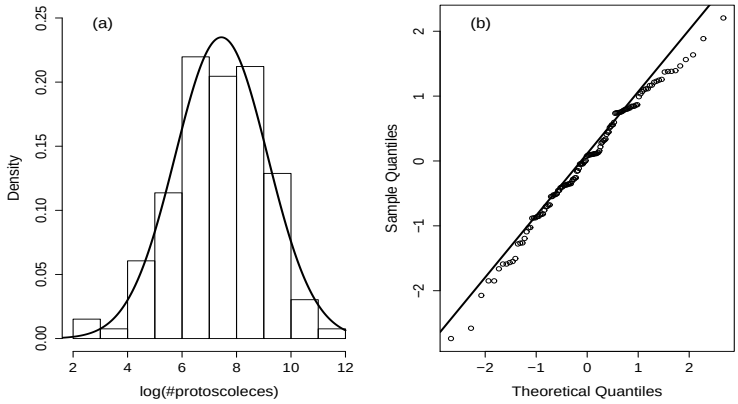

Figure 2: (a) *Histogram and estimated normal distribution and* (b) *QQ-plot of the log-transformed protoscolex burdens in fertile Echinococcus granulosus cysts.*

4. Combined model for the transmission dynamics

4.1. Interaction model

We now introduce the scheme of the interaction model for the whole life cycle of *Echinococcus granulosus*. Computational aspects for the simulation of the model are discussed in Section 5. A summary of the model parameters and a discussion on how we fix the parameters is given at the end of this Subsection.

Let the constant population sizes of dogs and sheep be $n^{(1)}$ and $n^{(2)}$. Transmission is assumed to take place in a homogeneous, homogeneously mixing closed community. Let the sub-models for the dog and sheep populations be given by the clumped infection processes presented in Section 2, with parameters fixed by their estimates as given in Table 3. The fertility of cysts is described by (6), with parameters k^*, a and b. Assume that dogs have exponentially distributed life times with mean

r, which is suggested by the dog sample from Kazakhstan. The sheep sample from Kazakhstan which contains the ages of sheep at death does not suggest any reasonable parametric distribution, and thus the distribution of ages at death is approximated by the empirical distribution in the sample. Suppose that infection severity does not influence the lifespan of sheep and dogs and that all animals at death are replaced by susceptibles (newborns) of the same type.

The infection contact scheme to link the two sub-models and the fertility model are connected as follows. A sheep dies (or is slaughtered) at age t, having a number of cysts n_c, specified by its infection history. The cadaver, ingested by a single dog, is considered to be infectious if at least one cyst is fertile. Thus the probability of infectiousness of the sheep is $1 - [1 - q(t)]^{n_c}$, where $q(t)$ is as in (6). If the sheep cadaver is infective, the dog eating it is infected with a lognormally $\mathrm{LN}(\mu, \sigma^2)$ distributed number of parasites as discussed in Section 2. The ingested parasite burden then develops according to the decay dynamics given by the PT respectively the CS model.

For the infection of sheep through excreta of infective dogs, we assume that all infected dogs are equally infectious, independent of their parasite burden. There are no experimental studies for *Echinococcus granulosus* available which accurately investigate the parasite burden in dogs and the resulting egg output. However, experiments with *Echinococcus multilocularis*, a comparable parasite, suggest that the egg output for a given parasite burden in dogs is highly variable. Eckert & Deplazes (2004) suggested that there may be density-dependent constraints reducing the fertility of parasites for larger burdens. These two points makes the above assumption somewhat reasonable. Let η be a random variable for the rate of potentially infectious contacts of the sheep with excreta of dogs. Let δ denote the prevalence of infection in dogs. Since only a fraction δ of the contacts of sheep are with excreta of infected dogs, the infection rate of sheep is $\eta\delta$. The sub-process for the sheep population (1) indicates that $\eta\delta$ has a gamma distribution with parameters ψ and ξ. Hence the contact rate η has a gamma distribution with the same shape parameter ψ, but different scale parameter $\tilde{\xi} = \xi/\delta$. In the model, a sheep obtains an individual contact rate η at birth, so that the resulting infection rate at time t is $\eta\delta$, specifying its potential to get infected.

A summary of all parameters of the model and their fixed values is given in Table 3. The parameters for the sub-process of the dog population are fixed by their estimates given in Table 1. For the sheep sub-process, the parameters ψ, θ and ζ are fixed by their estimates given in Table 1, and $\tilde{\xi}$ is fixed as $\hat{\xi}/0.230 = 0.343/0.230 = 1.491$ by noting that 0.230 is the prevalence of infection in dogs of the Kazakhstan sample. The parameters for the fertility model are fixed by their estimates from the Kyrgyzstan sample, computed in the previous Section. Finally, the remaining model parameters are fixed as follows. Assuming that for the dog sample of Kazakhstan, approximately every fifth dog was sampled (rough estimate of participants of that study), we set the constant population size of dogs $n^{(1)} = 3030$, which corresponds

to 5 times the sample size of the dog sample from Kazakhstan. The population ratio ρ can be approximated by 10.7, based on field data in Kazakhstan, where $\rho = 10.368$ $(95\%CI : 10.074, 10.706)$ (sample from 1 village; unpublished data), and in Kyrgyzstan, where $\rho = 11.418$ $(95\%CI : 10.593, 12.383)$ (samples from 3 villages; unpublished data), and where during a purgation study in dogs, owners were asked how many sheep and dogs they own. Hence the constant population size of sheep can be set to $n^{(2)} = \rho n^{(1)} = 32421$. And finally, r is fixed by the estimated mean age of 3 years from the dog sample from Kazakhstan.

Table 3: Summary of all parameters with fixed values needed to simulate the mechanistic two-host model. Note that " − " for the dog models indicates that the corresponding parameter is not specified in that model. The infection rate β is only used for the initialization of the dog population.

Dogs	PT	CS	Explanation
β	0.445	0.340	Parameter PT/CS sub-processes for dogs
μ	6.001	4.302	" "
σ	2.955	2.616	" "
λ	8.833	−	Parameter PT sub-process for dogs
t_d	−	0.744	Parameter CS sub-process for dogs
Sheep			
ψ	0.941		Parameter for gamma mixture of contact rates
$\tilde{\xi}$	1.491		" "
θ	0.351		Parameter of clump size distribution
ζ	5.859		" "
Fertile			
k^*	0.103		Parameter in fertility model (6)
a	0.124		" "
b	1.394		" "
Model			
$n^{(1)}$	3030		Population size dogs
ρ	10.7		$\rho = n^{(2)}/n^{(1)}$, so that population size sheep $n^{(2)} = 32421$
r	3		Sample mean age of dogs

4.2. Basic reproduction number

Assume that we have a single infected animal in an otherwise fully susceptible dog population of size $n^{(1)}$, and that all $n^{(2)}$ sheep are uninfected and susceptible. The infected dog infects sheep with cysts at a mean rate of $\varphi := \psi\tilde{\xi}\rho$, where $\rho = n^{(2)}/n^{(1)}$. Given that a sheep at death has at least one fertile cysts, it transmits the disease to exactly one dog. Let ω denote the mean proportion of sheep that are infectious

conditional on having cysts. Given model (6) for the fertility of a cyst in a sheep of age t and model (1) describing the cyst load of a sheep of age t, the mean fertility of a sheep of age t having cysts becomes $\omega_t = \sum_{j=1}^{\infty} \mathbb{P}(Y_t = j)[1 - (1 - q(t))^j]$, where $\mathbb{P}(Y_t = j)$ is specified by $\mathcal{P}_t$ in (1). If $t_1, \ldots, t_n$ are the ages of the sheep in the sample, then ω can be defined as $\omega := (1/n) \sum_{l=1}^{n} \omega_{t_l}$.

Hence on average $\varphi\omega$ infections will arise in the dog population, per unit duration of the dog's infectious period. Let α_1 be the mean loss rate of infection in dogs, which is the sum of the natural death rate of dogs and a rate of loss of infection through death of parasites. For the PT model, we have seen that the mean survival time of a single infection is given as $\mathbb{E}(T) = (\gamma + \mu)/\lambda$ with γ the Euler-Mascheroni constant, and for the CS model, the duration is fixed with t_d. Given that the dogs have an exponential life time with mean 3 years, we have $\alpha_1 = 1/3 + \lambda/(\gamma + \mu)$ for the PT model and $\alpha_1 = 1/3 + 1/t_d$ for the CS model. Then a single infected dog indirectly infects on average a total of $\varphi\omega/\alpha_1$ dogs if all sheep are susceptible.

Theorem 4.1. *The basic reproduction number*

$$R_1 := \frac{\varphi\omega}{\alpha_1},$$

with parameters as above, is a threshold such that, as $t \to \infty$, $R_1 < 1$ implies that the disease dies out and $R_1 > 1$ implies that the disease persists.

The threshold R_1 can be computed as follows. Fix ψ, $\tilde{\xi}$ and ρ as in Table 3. As seen in Section 2, the mean survival time of a single infection in dogs is comparable in the PT and CS models, with a common value of about 0.75 years. Thus α_1 is fixed to $1/3 + 1/0.75 \approx 1.65$. And finally, ω is fixed as follows. Let the model parameters for (6) and (1) be fixed by their estimates as before. For each age t in the Kazakhstan sample, we can approximate ω_t by $\omega_t(m) := \sum_{j=1}^{m} \mathbb{P}(Y_t = j)[1 - (1 - q(t))^j]$, where m is chosen such that $\omega_t(m) - \omega_t(m-1) < 10^{-8}$. Given the ages of the sheep $t_1, \ldots, t_n$, we have $\omega = (1/n) \sum_{l=1}^{n} \omega_{t_l}(m_l)$, where m_l is the value of m computed for age t_l. Here, $\omega = 0.198$, which is close to the corresponding empirical value of 0.205, where instead of the theoretical distribution for the number of cysts the observed distribution is used. Thus we obtain $R_1 = 1.806$.

As a marginal note, subdividing the ages of the sheep sample into age classes $(0, 1]$, $(1, 2]$, $(2, 3]$, $(3, 4]$, $(4, 5]$ and $5+$, where the class $5+$ summarizes all sheep older than 5 years, and computing ω, the mean infectiousness conditional on having cysts, for all of the classes separately results in 0.051, 0.155, 0.268, 0.354, 0.414 and 0.455. This indicates that ω restricted to young sheep ≤ 1 year is about 5%, whereas it is about 45% in sheep older than 5 years.

5. Application

5.1. Tau-leaping

Simulation experiments are carried out to investigate the properties of the interaction model. We use the tau-leaping method (Gillespie 2001), an approximation to the Gillespie algorithm (Gillespie 1977) to speed up simulation. Given a subinterval of length τ, the expected number of death and infection events assuming constant rates over the interval is determined and executed and the state variables of the system such as ages and loads are updated. The choice $\tau = 0.01$ is reasonable since during τ, the net change of δ, the prevalence of infection in dogs, is about $\tau|\beta - \alpha_1|$, with β the infection rate of dogs approximated by 0.4 per dog per year as mean value of the infection rates β of the PT and CS model given in Table 1, and α_1 approximated by 1.65 as before. Thus the net change is about 1.2%, ensuring that the infection pressure towards sheep does not change greatly during τ. In addition, the mean loads of dogs will decrease during τ on average by a factor of $\exp(-0.01\lambda) = 0.915$, where λ is fixed by its estimate in Table 1. This is sufficiently accurate for our application since all dogs harboring parasites are considered to be equally infectious.

Let the model parameters be fixed as given in Table 3. Starting with $n^{(1)}$ dogs and $n^{(2)}$ sheep, the host populations are initialized by attributing an actual age with corresponding load and a remaining life duration to all animals. For each dog, a lifespan is generated by using an exponential distribution with mean r years and the corresponding load is computed based on the PT or CS processes. To obtain an appropriate starting distribution for the ages of the sheep population, we note that the age at death distribution of sheep is length-biased (Simon 1980). The lifetime L is thus sampled from the empirical age at death distribution, weighted in proportion to lifetime. The age of the sheep is determined by a realization of a $U[0, L]$-distribution, and a load for the sheep is then computed based on the process (1). This provides a satisfactory initialization so that equilibrium can be reached reasonably quickly.

After each time step τ, the number of death events in both host populations is computed and the dead animals are removed and replaced by uninfected newborns with age 0. To each newborn sheep, a contact rate η with excreta of dogs is drawn from a gamma distribution with shape parameter ψ and scale parameter $\tilde{\xi}$ as seen in the previous Section. The infection rate towards a sheep at time t with prevalence of infection in dogs $\delta(t)$ is then $\eta\delta(t)$. The number k of infections in the dog population is then determined as described in Subsection 4.1, and k dogs are randomly selected from the $n^{(1)}$ dogs in the population. A realization of $\mathrm{LN}(\mu, \sigma^2)$ is attributed to each of those k dogs. The ingested parasite load then survives for a fixed time t_d in the CS model. In the PT model, a dog having at time t j parasites will lose on average $j(1 - \exp(-\lambda\tau))$ during the next time step τ. Thus for the simulation of the PT model, we assume that parasites die independently with probability $1 - \exp(-\lambda\tau)$ during the time step τ.

The prevalence of infection in dogs $\delta(t)$ is evaluated, and the number of infectious

events per sheep with contact rate η during the time step τ is given as a realization l of $\mathrm{Po}(\eta\delta(t)\tau)$. The sheep is then infected l times with a number of cysts specified by a realization of a zero-truncated version of a negative binomial random variable with shape parameter θ and scale parameter ζ. Since τ is sufficient small, there will rarely be more than one infection per sheep per time step. Finally, the ages of the sheep which did not die are updated, and the ages and the parasite burden in dogs which did not die are updated, where the update of the burden in dogs depends on whether the PT or CS model is used.

For the simulation, a suitable burn-in period for the system to reach its endemic equilibrium state is determined such that the relative change of the mean of any of the quantities of interest below averaged over consecutive intervals of size 500 time steps τ (which corresponds to 5 years simulation time) is less than 0.005. For the present application, the burn-in period is fixed at 20000 steps.

5.2. Results

The quantities investigated in this simulation study for the host populations are prevalences of infection δ in dogs and s in sheep, the per capita contact rates $\kappa^{(1)}$ of dogs with sheep viscera and the per capita contact rate $\kappa^{(2)}$ of sheep with excreta from dogs. The contact rate $\kappa^{(1)}$ of dogs is computed at each time step τ in the simulation model as the total number of sheep dying divided by $n^{(1)}\tau$. Similarly, $\kappa^{(2)}$ is computed at each time step τ as the total number of contacts of sheep with excreta of dogs divided by $n^{(2)}\tau$. Hence the above quantities can be determined at each time step τ in the model. In addition, we sample their values at every 2000th time unit τ after the burn-in period until we have 1000 values. The 2.5% and 97.5% quantiles can be computed to obtain an idea about the variation of the quantities of interest.

The results can then be compared to their corresponding estimates derived from the sheep and dog samples from Kazakhstan. In what follows, we refer to these estimates as the true values. The true values of the prevalences of infection δ and s are computed from the corresponding samples, resulting in in 0.230 for dogs and 0.363 for sheep. The true value of $\kappa^{(2)}$ is given by $\psi\tilde{\xi}$, where $\tilde{\xi} = \xi/0.230$, and ψ and ξ are fixed by their estimates given in Table 3. Finally, the true value of the contact rate $\kappa^{(1)}$ of dogs is given by $\beta/s\omega$, where β is the infection rate β for the PT respectively the CS model as discussed in Section 2, and ω is the average probability that a sheep harboring cysts is infectious, as discussed in the previous Section and computed as 0.198. Hence using the estimates of β given in Table 3, it follows that $\kappa^{(1)}$ is 6.191 in the PT setting and 4.731 in the CS setting.

Table 4 shows the above quantities of interest of the interaction model with the PT respectively CS sub-model for the dog population, together with the true values introduced above. The values of the interaction model with the PT sub-process are in line with the true values. The computed contact rate $\kappa^{(1)}$ of dogs with viscera from sheep in the simulation model is with 5.986 close to the true value 6.191 in

Table 4: Prevalences of infection δ in dogs and s in sheep, and the contact rates $\kappa^{(1)}$ of dogs with viscera from sheep, and $\kappa^{(2)}$ of sheep with excreta from dogs, obtained in the simulation model, together with the true values, as described in Subsection 5.2. Note that there are two true values of $\kappa^{(1)}$ for either the PT (value=6.191) or the CS model (value=4.731).

	TRUE	Simulation/PT	Simulation/CS
δ	0.230	0.232 (0.211, 0.255)	0.252 (0.235, 0.274)
s	0.363	0.346 (0.322, 0.368)	0.366 (0.341, 0.379)
$\kappa^{(1)}$	6.191/4.731	5.986 (3.408, 9.737)	6.099 (3.833, 10.127)
$\kappa^{(2)}$	1.348	1.384 (1.096, 1.739)	1.277 (1.069, 1.684)

Table 1. In the model with the CS sub-process, the prevalence of infection in the dog population is not well reflected, and the contact rate $\kappa^{(1)}$ of 6.099 is much larger than the corresponding true value 4.731 in Table 1. Thus we will use the interaction model with the PT sub-process for the dog population for investigating the influence of environmental factors and control interventions on the dynamics of the life-cycle of *Echinococcus granulosus*. Figure 3 shows the positive cyst burdens in sheep for different age classes, obtained from a snapshot of the simulation model after the burn-in period, together with the observed positive burdens in the sheep sample from Kazakhstan. For simplification, we represent the positive cyst counts from the simulation output through a kernel density estimator. Analogously, Figure 4 represents a snapshot for the log-transformed positive counts of dogs together with a histogram of the corresponding observed quantities from the dog sample from Kazakhstan. The outputs are in line with the observed counts for sheep and dogs.

5.3. Environmental influence: Seasonality

Eggs released by dogs are subject to environmental effects. Climate and temperature are density-independent constraints limiting the survival of eggs (Thompson & Lymbery 1986). Infectious eggs have been found in water and damp sand for 20 days at $30°C$, 32 days at $10-21°C$. and 225 days at $6°C$ (Thompson & Lymbery 1986). The eggs survive in general for only short periods if they are exposed to direct sunlight and dry conditions. Sweatman & Williams (1963) have showed that the survival of infectious *Echinococcus granulosus* eggs can reach up to 41 months in nature under varying temperatures ranging from $-3°C$ to $38°C$. Veit et al. (1995) suggest that the survival for *Echinococcus multilocularis* eggs, which are comparable to *Echinococcus granulosus* eggs, is of the order of 100 days. In Kazakhstan, the maximum average monthly temperature is accounted in July at $25°C$ and the lowest in January at $-6°C$. In Middle June to late August, the temperatures are approximately $20°C$ and there is a low rainfall with an average of $35mm$.

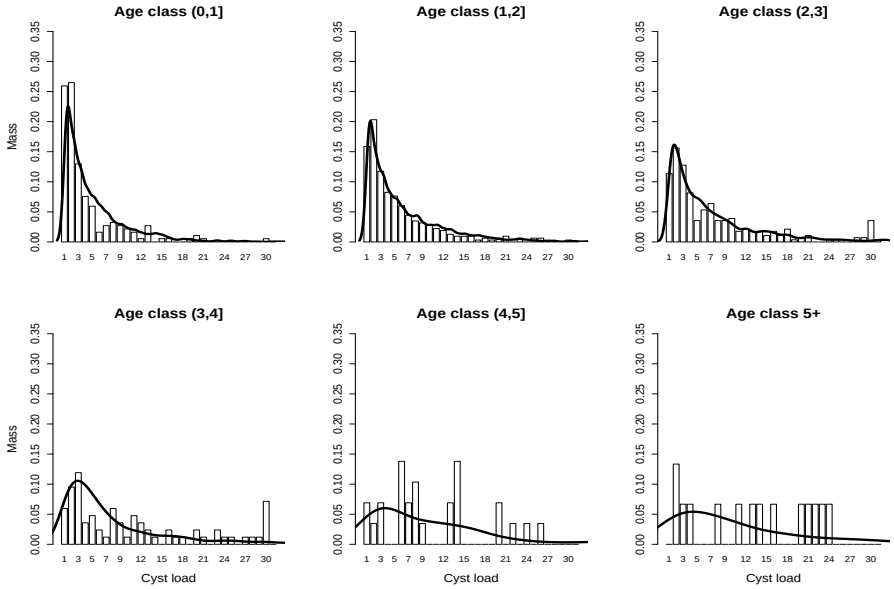

Figure 3: *Snapshot of the distribution of positive burdens of Echinococcus granulosus cysts in sheep from the simulation model (solid curve; computed by a kernel density estimation) with a histogram of the observed positive burdens of the sheep sample from Kazakhstan for different age classes. The age classification is taken from Heinzmann et al. (2009), with* 5+ *the age class summarizing all sheep older than* 5 *years.*

This seasonality is implemented in the interaction model as follows. Given the individual contact rates η for each sheep, its infection rate is defined as $\eta\delta$, where δ is the prevalence of infection in dogs. The survival time of eggs in the 2.5 months from Middle June to late August is approximately 3 times (32 days versus 100) smaller than in the rest of the year. Hence we set the infection rate for a sheep equal to $\eta\delta r/3$ for Middle June to late August and equal to $\eta\delta r$ for the remaining year, where r is chosen such that $(9.5[\eta\delta r] + 2.5[\eta\delta r/3])/12 = \eta\delta$, indicating that the mean infection rate over the year of that sheep is still $\eta\delta$. This yields $r = 1.161$.

For the simulation with the implemented seasonality effect, the quantities of interest are computed as before, yielding prevalences of infection of $\delta = 0.233$ $(0.211, 0.256)$ in dogs and $s = 0.349$ $(0.326, 0.371)$ in sheep. The contact rates are $\kappa^{(1)} = 6.102$ $(3.632, 9.825)$ and $\kappa^{(2)} = 1.401$ $(1.124, 1.773)$. These are all close to the simulation values as above. This indicates that seasonality does not greatly influence the transmission dynamics, the sheep population acting as buffer in the transmission.

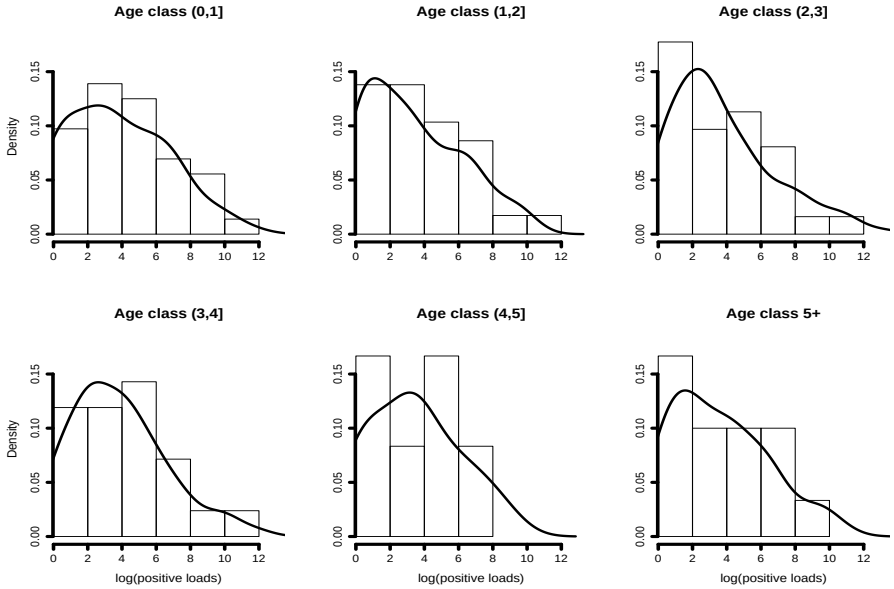

Figure 4: *Snapshot of the distribution of the log-transformed positive parasite loads in dogs from the simulation model (solid curve; computed by a kernel density estimation) with a histogram of the observed loads of the dog sample from Kazakhstan for different age classes, analolously to Figure 3 for sheep.*

5.4. Control interventions in dog population

To develop and evaluate public health control interventions against *Echinococcus granulosus*, it is imperative to understand the reaction of the transmission system to interventions. Mass dog treatment programs are widely applied in practice to control or eradicate *Echinococcus granulosus* (Gemmell 1958, Gemmell et al. 1958, Thompson & Lymbery 1986, Torgerson & Heath 2003d).

Based on our model, two scenarios are tested, both are based on a treatment of a certain proportion of dogs with an anti-parasitic drug every 6 weeks. The 6-week interval is based on the prepatent period of infection with *Echinococcus granulosus* and is the suggested treatment frequency for such control interventions (Cabrera et al. 2002). It is assumed that the drug eliminates the disease. In Kazakhstan, dogs stay mostly around households (discussion with study participants). Thus in scenario 1, we assume that 75% of dogs from the whole population are randomly selected and treated at each intervention. In scenario 2, we increase the percentage to 95%, reflecting a larger control effort. The distribution of the time to extinction is then approximated by simulation. The treatment is started after the initial burn-in period. For scenario 1, the mean time to extinction of the disease is 13.6 years (95%CI:11.2, 16.3), whereas for the scenario 2, it is 11.7 years (10.2, 14.5).

The large values of the mean extinction time are due to the fact that the mean

generation time for the cycle of infection in dogs is several years long, because the parasite has to wait in a sheep until it dies. The control interventions considered have the effect of reducing the mean duration $1/\alpha_1$ of infection, where α_1 is the sum of the death rate of dogs $1/r = 1/3$ and the inverse of the mean duration of an infection. For both scenarios, given that a dog is infected at the beginning of treatment k, he will lose its infection at the $(k+i)$th treatment $(i = 0, 1, \ldots)$ according to a geometric distribution with probability $p = 0.75$ respectively $p = 0.95$. Since the mean of the geometric distribution is $(1-p)/p$, the dog will stay infected for a mean time of $(1-p)/p \cdot 6$weeks, resulting in 2 weeks for scenario 1 and 0.3 for scenario 2. The infection time point of the dog is uniformly distributed between the $(k-1)$th and kth treatment, thus on average 3 weeks prior to its first treatment. Hence the mean duration of an infection becomes 5 weeks for scenario 1 and 3.3 weeks for scenario 2. We have seen in Subsection 4.2 that $R_1 = 2.980/\alpha_1$. Hence $R_1 \approx 0.3$ for scenario 1 and $R_1 \approx 0.2$ in scenario 2. Let the generation time be the time of infection of a dog until it infects indirectly another dog. Hence, starting from around $0.230n^{(1)} \approx 700$ infected dogs, scenario 1 implies that $5-6$ generations are needed to eliminate infection since $700(R_1)^6 < 1$, and in scenario 2, $4-5$ generations are needed. This suggests that, with the lifetime distribution observed in the sheep population, the mean generation time of an infection is around 2.5 years.

If a prepatent period of infection were included in the model, a rather smaller number of generations would be needed to eliminate infection.

6. Discussion

In this paper, an interaction model is proposed to describe the life-cycle of the parasite *Echinococcus granulosus* in its two-host system between dogs and sheep. The model architecture is based on compound processes for the sheep population (Heinzmann et al. 2009), shot noise processes with absorption mechanisms at zero for the dog population (Heinzmann et al. n.d.) and a biologically reasonable contact scheme for the inter-population infections. The fertility of cysts in sheep and thus the infectiousness towards dogs is represented by a two-part conditional model, fitted to field data. The results indicate that cysts of age 2 have an average probability of 6% to be fertile and the asymptotic probability of fertility of a cyst is 10%, indicating that 1 out of 10 older cysts are fertile. Furthermore, the mean infectiousness of a sheep conditional on harboring cysts is about 20%. In sheep of age 1 or younger, the mean infectiousness is about 5%, whereas in sheep older than 5 years, it is about 45%. The mean infectiousness in older sheep is higher since they have on average more and older cysts than younger sheep.

A threshold of extinction R_1 is derived for the interaction model, such that if the time $t \to \infty$, $R_1 > 1$ indicates persistence of the disease. We show that for our data from Central Asia, $R_1 \approx 1.8$ is plausible. The interaction model is then investigated by simulation. It is shown that the model with exponential decay dynamics of the

worm burden in dogs performs better than that with a constant duration of infection in dogs. Finally, the sensitivity of the interaction model towards environmental factors and control interventions is investigated. It is shown that the model output for seasonally varying infectiousness of excreta from dogs is close to the output of a model with the same average infectiousness, but held constant throughout the year, indicating that the sheep population acts as a buffer on seasonally varying environmental influences. Different mass treatment schemes are tested and it is shown that the infection can persist in the population for many years despite large control efforts.

Achnowledgements The work was supported by the Schweizerischer Nationalfonds (SNF), project no. 107726.

References

Aminzhanov, M. (1975), 'Duration of the life of Echinococcus granulosus in the organism of dogs', *Veterinariia* **12**, 70–72.

Bass, F. (1969), 'A new product growth model for consumer durables', *Management Science* **15**, 215–227.

Cabrera, P. A., Haran, G., Benavidez, U., Valledor, S., Perera, G., Lloyd, S., Gemmell, M. A., Baraibar, M., Morana, A., Maissonave, J. & Carballo, M. (1995), 'Transmission dynamics of Echinococcus granulosus, Taenia hydatigena and Taenia ovis in sheep in Uruguay', *Int J Parasitol* **25**, 807–813.

Cabrera, P. A., Lloyd, S., Haran, G., Pineyro, L., Parietti, S., Gemmell, M. A., Correa, O., Morana, A. & Valledor, S. (2002), 'Control of Echinococcus granulosus in Uruguay: evaluation of different treatment intervals for dogs', *Vet Parasitol* **103**, 333–340.

Cohen, A. C. (1960), 'An extension of a truncated Poisson distribution', *Biometrics* **16**, 447–450.

Duan, N., Manning, W. G. J., Morris, C. & Newhouse, J. (1984), 'Choosing between the sample selection model and the multi-part model', *JBES* **2**, 283–289.

Eckert, J. & Deplazes, P. (2004), 'Biological, epidemiological and clinical aspects of Echinococcosis, a zoonosis of increasing concern', *Clin Microbiol Rev.* **17**, 107–135.

Economides, P. & Cristofi, G. (2002), *Cestode zoonoses: Echinococcosis and cysticercosis. An emergent and global problem*, 3 edn, NATO Science Series:IOS Press Amsterdam.

Gemmell, M. A. (1958), 'Hydatid disease in Australia, III. Observations on the incidence and geographical distribution of hydatidiasis in sheep in New South Wales', *Aust Vet J* **34**, 269–280.

Gemmell, M. A. (1959), 'Hydatid diseases in Australia. IV. Observations on the incidence of Echinococcus granulosus on stations and farms in endemic regions of New South Wales', *Aust Vet J* **35**, 396–402.

Gemmell, M. A., Lawson, J. R. & Roberts, M. G. (1986), 'Population dynamics in echinococcosis and cysticercosis: biological parameters of Echinococcus granulosus in dogs and sheep', *Parasitology* **92**, 599–620.

Gemmell, M. A., Oudemans, G. & Sakamoto, T. (1958), 'The effect of bithionol sulphoxide on Echinococcus granulosus and Taenia hydatigena infections in dogs', *Res Vet Sci* **18**, 109–110.

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', *J Chem Phys* **81**, 2340–2361.

Heinzmann, D., Barbour, A. D. & Torgerson, P. R. (2009), 'Compound processes as models for clumped parasite data'. accepted, Math. Biosci.

Heinzmann, D., Barbour, A. D. & Torgerson, P. R. (n.d.), 'Shot noise processes for clumped infections with time-dependent decay dynamics'. submitted.

Roberts, M. G., Lawson, J. R. & Gemmell, M. A. (1986), 'Population dynamics in echinococcosis and cysticercosis: Mathematical model of the life-cycle of Echninococcus granulosus', *Parasitology* **92**, 621–641.

Simon, R. (1980), 'Length-biased sampling in ethiological studies', *Am. J. Epidem.* **111**, 444–452.

Sweatman, G. K. & Williams, R. J. (1963), 'Survival of echinococcus granulosus and taenia hydatigena eggs in two extreme climatic regions of new zealand', *Res. Vet. Sci.* **4**, 199–216.

Thompson, R. C. A. & Lymbery, A. J. (1986), *The biology of Echinococcus and hydatid disease*, London: George Allen and Unwin.

Todorov, B. & Boeva, V. (1999), 'Human echinococcosis in Bulgaria: a comparative epidemiologiocal analysis', *Bulletin WHO* **77**, 110–118.

Torgerson, P. R., Burtisurnov, K. K., Shaikenov, B. S., Rysmukhambetova, A. T., Abdybekova, A. M. & Ussenbayev, A. E. (2003a), 'Modelling the transmission dynamics of Echinococcus granulosus in dogs in rural Kazakhstan', *Parasitology* **126**, 417–424.

Torgerson, P. R. & Heath, D. D. (2003d), 'Transmission dynamics and control options for cystic echinococcosis', *Parasitol.* **127**, 143–158.

Torgerson, P. R., Oguljahan, B., Muminov, M. E., Karaeva, R. R., Kuttubaev, O. T., Aminjanov, M. & Shaikenov, B. (2006), 'Present situation of cystic echinococcosis in Central Asia', *Parasitol Int.* **55**, 207–212.

Torgerson, P. R., Shaikenov, B. S., Rysmukhambetova, A. T., Ussenbayev, A. E., Abdybekova, A. M. & Burtisurnov, K. K. (2003b), 'Modelling the transmission dynamics of Echinococcus granulosus in sheep and cattle in Kazakhstan', *Vet Parasitol* **114**, 143–153.

Torgerson, P. R., Williams, D. H. & Abo-Shehada, M. N. (1998), 'Modelling the prevalence of Echinococcus and Taenia species in small ruminants of different ages in northern Jordan', *Vet Parasitol* **79**, 35–51.

Veit, P., Bliger, B., Schad, V., Schaefer, J., Frank, W. & Lucius, R. (1995), 'Influence of environmental factors in the infectivity of Echinococcus multilocularis eggs', *Parasitology* **110**, 79–86.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. (1996), 'Modeling the abundance of rare animals: statistical models for counts with extra zeros', *Ecological Modeling* **80**, 297–308.

Ziadinova, I., Mathis, A., Trachsel, D., Rysmukhambetova, A., Abdyjaparov, T. A., Kuttubaev, O., Deplazes, P. & Torgerson, P. R. (2008), 'Canine echinococcosis in Kyrgyzstan: Using prevalence data adjusted for measurement error to develop transmission dynamics models', *Int J Parasitol.* **38**, 1179–1190.

Coupling of an epidemic model to a branching process: Introduction

This chapter describes the motivation which led to the fourth paper of the dissertation. Starting with a prevalence-based epidemic model for *Echinococcus granulosus*, we wanted to determine its time to extinction. For this purpose, we coupled the model to a multitype Markov branching process, for which we then derived the time to extinction in the fourth paper.

First, the epidemic model and its approximating branching process are introduced. Then we couple the two models and show that they coincide if the number of susceptibles is large as compared to the number of births (new infections) in the processes. Thus the result of the fourth paper can be used to approximate the time to extinction in the epidemic model, by way of its approximating branching process. The satisfactory performance of the approach is illustrated in the sequel of the following paper. To distinguish the labelling from that of the fourth paper, we use a preceding "A" for section, equation and theorem numbers. All references for this chapter are listed at the end of the dissertation.

A.1. Prevalence-based model

Based on the natural life-cycle of *Echinococcus granulosus* (Eckert & Deplazes (2004)), we introduce an interaction model for the transmission of infection between dogs and sheep, the primary definitive and intermediate hosts. Suppose that transmission takes place in a homogeneous, homogeneously mixing closed community with constant population sizes of $n^{(1)}$ dogs and $n^{(2)}$ sheep. Let $\boldsymbol{E} = (D, S) = \{(D(t), S(t))\}_{t\geq 0}$ be the numbers of infective dogs and sheep at time t. The epidemic can be described as follows. Infective dogs infect susceptible sheep by indirect transmission based on free-living stages in their excreta. The contacts of individual sheep with the excreta of dogs is assumed to occur according to independent Poisson processes with rate θ. The rate θ depends on the density of infective dogs and the grazing activity of sheep, so that infection of a susceptible sheep occurs at rate $\theta D/n^{(1)}$. Infections are assumed to be permanent (Gemmell et al. 1986, Torgerson et al. 1998). Sheep live for an exponentially distributed time with rate λ_2 before they die (or are slaughtered) and are fed directly to a dog. An infection is established if the dog is susceptible and the dead sheep is infectious. The infectious period in dogs is exponentially distributed with rate λ_1 and the loss of infection happens either through loss of parasites or through death. It is further assumed that there is no acquired immunity (Gemmell et al. 1986, Torgerson et al. 2003a) and that all subjects at death are replaced by susceptibles (newborn) of the same type.

The process $\boldsymbol{E}$ takes values in $\{0,1,\ldots,n^{(1)}\}\times\{0,1,\ldots,n^{(2)}\}$ and is characterized by the following set of Markov transition rates:

Transition	Rate
$D \to D-1,\ S \to S$	$\lambda_1 D$
$D \to D,\quad S \to S+1$	$\theta(n^{(2)}-S)(D/n^{(1)})$
$D \to D,\quad S \to S-1$	$\lambda_2 S(D/n^{(1)})$
$D \to D+1,\ S \to S-1$	$\lambda_2 S(1-(D/n^{(1)}))$.

(A.1)

A.2. Approximating branching processes

Let $\boldsymbol{Z} = (Z_1, Z_2) = \{(Z_1(t), Z_2(t))\}_{t\geq 0}$ be a multitype Markov branching process, where Z_1 and Z_2 denote the number of animals of type 1 and 2 respectively, with corresponding transitions

Transition	Rate
$Z_1 \to Z_1-1,\quad Z_2 \to Z_2$	$\lambda_1 Z_1$
$Z_1 \to Z_1,\quad\ Z_2 \to Z_2+1$	$\theta\rho Z_1$
$Z_1 \to Z_1+1,\quad Z_2 \to Z_2-1$	$\lambda_2 Z_2$.

(A.2)

The process (A.2) represents a birth and death process, with events (i) an animal of type 2 lives for an exponential time of rate λ_2 and produces at its death one offspring of type 1, (ii) an animal of type 1 lives for an exponential time with rate $\lambda_1+\theta\rho$ and produces at its death either no offspring with probability $\lambda_1/(\lambda_1+\theta\rho)$ or one type 1 and one type 2 offspring with probability $\theta\rho/(\lambda_1+\theta\rho)$, where $\rho = n^{(2)}/n^{(1)}$ is the population ratio.

Let $z_1 := Z_1/n^{(1)}$ and $z_2 := Z_2/n^{(2)}$. Then the mean field dynamics are given by

$$\frac{dz_1}{dt} = -\lambda_1 z_1 + \rho\lambda_2 z_2\,,$$
$$\frac{dz_2}{dt} = \theta z_1 - \lambda_2 z_2\,.$$

Applying the results given in Diekmann et al. (1990) and Heesterbeek & Roberts (2007), it is straightforward to verify that he type-reproduction number R_1, a threshold for the extinction of the process, is given by the following result.

Theorem A.1. *The quantity*

$$R_1 := \frac{\theta\rho}{\lambda_1}$$

is a threshold for the deterministic model above such that as $t\to\infty$, $R_1<1$ implies that $(z_1,z_2)\to(0,0)$ and $R_1>1$ implies that there $(z_1,z_2)\to(\bar{z}_1,\bar{z}_2)$, where

$$\bar{z}_1 = \frac{\lambda_2(\rho\theta-\lambda_1)}{\theta(\rho\lambda_2+\lambda_1)} \quad \textit{and} \quad \bar{z}_2 = \frac{\rho\theta-\lambda_1}{\rho(\theta+\lambda_2)}\,.$$

We will see that the epidemic process $\boldsymbol{E}$ and the branching process $\boldsymbol{Z}$ can be constructed on a same probability space so that there is a direct correspondence between the number of infective dogs D and the number of type 1 individuals Z_1, respectively between the number of infective sheep S and the number of type 2 animals Z_2. It is shown that the construction implies that $D \leq Z_1$ and $S \leq Z_2$ almost surely. Hence $R_1 < 1$ for the branching process implies extinction behavior in $\boldsymbol{E}$.

Under some assumptions that we will discuss below, the construction of the processes on a same probability space indicates that $\boldsymbol{Z}$ and $\boldsymbol{E}$ coincide with high probability. Then, the biological interpretation of R_1 is as follows. The mean duration of an infection in dogs is $1/\lambda_1$. Given an infectious dog, it infects sheep at rate $\theta\varrho$. Thus the expected number of sheep infected by a single infectious dog is R_1. Since an infected sheep is connected with exactly one dog, R_1 is the mean number of infections in the dog population caused (indirectly) by a single infectious dog.

A.3. Coupling

Assume that $R_1 < 1$, so that the processes (A.2) and thus (A.1) are sub-critical as seen before. Let $\boldsymbol{I} = (I_1, I_2)$ be the initial numbers of infective dogs and sheep respectively, and denote with $\boldsymbol{M} = (M_1, M_2)$ the initial numbers of susceptible dogs and sheep respectively so that $M_i = n^{(i)} - I_i$ $(i = 1, 2)$. Denote by $\boldsymbol{E}_{\boldsymbol{I}}^{\boldsymbol{M}}$ the epidemic process (A.1) and by $\boldsymbol{Z}_{\boldsymbol{I}}$ the (approximating) branching process (A.2). Note that both processes are Markov.

We use the construction argument of Ball (1983) and Ball & Donnelly (1995) to couple $\boldsymbol{E}_{\boldsymbol{I}}^{\boldsymbol{M}}$ and $\boldsymbol{Z}_{\boldsymbol{I}}$. They described the construction of a single-host epidemic model from a limiting branching process. They showed that if the branching process is sub-critical, the epidemic and branching processes coincide for $N \to \infty$, where N is the number of susceptible hosts. For that, we need to adapt to our model the independent and identically distributed life histories of the individuals, given as (L, ξ) in Ball & Donnelly (1995), where L is the time elapsing between an individual's infection and its death, and ξ is a point process of times at which contacts are made. We specify the life histories for dogs as (L_1, ξ_1), where L_1 is exponentially distributed with rate λ_1 and ξ_1 is a point process of rate $\theta\rho$ at which sheep make infective contacts with its excreta, and the life histories for sheep with (L_2, ξ_2), where L_2 is exponentially distributed with rate λ_2 and $\xi_2[0, L_2) = 0$ and $\xi_2\{L_2\} = 1$, since an infected sheep is connected with exactly one dog and the infection is transmitted at death of the sheep. The construction of the process is now similar to the construction in the proof of Theorem 2.1 in Ball & Donnelly (1995), except that in our case, individuals contacted during an infection event are chosen independently and uniformly from the M_i $(i = 1, 2)$ initial susceptibles in the corresponding host population. It follows that $D \leq Z_1$ and $S \leq Z_2$ almost surely.

Let B_1 and B_2 be the random variables for the total number of new births of type 1 and 2 individuals respectively into the branching process (A.2).

Lemma A.3. *We have*

$$\mathbb{E}(B_1|\boldsymbol{I} = (I_1, I_2)) = 2a(I_1 + I_2) + I_2\,,$$
$$\mathbb{E}(B_2|\boldsymbol{I} = (I_1, I_2)) = a(I_1 + I_2)\,,$$

where $a = \theta\rho/(\lambda_1 - \theta\rho)$.

Proof. Define $m_i := \mathbb{E}(B_i|\boldsymbol{I} = (1,0))$ and $k_i := \mathbb{E}(B_i|\boldsymbol{I} = (0,1))$ for $i = 1, 2$, where $\boldsymbol{I} = (1,0)$ highlights that the branching process is started with a single type 1 individual and $\boldsymbol{I} = (0,1)$ analogously. Define $a := \theta\varrho/(\lambda_1 - \theta\varrho)$. Starting with a type 1 individual, we can have a splitting into a type 1 and type 2 individual with probability $p := \theta\varrho/(\lambda_1 + \theta\varrho)$, or no offspring with probability $1 - p$. When starting with a type 2 individual, there will be exactly one offspring of type 1, thus $m_1 = p(1 + m_1 + k_1)$ and $m_2 = p(1 + m_2 + k_2)$. We have $k_1 = 1 + m_1$ and $k_2 = m_2$. Since $R_1 < 1$, then m_i and k_i, for $i = 1, 2$, are finite. Then, using $k_1 = 1 + m_1$ in the expression for m_1 implies that $m_1 = 2a$ and thus $k_1 = 2a + 1$. Analogously, we obtain $m_2 = k_2 = a$. Hence the lemma follows immediately. □

Lemma A.4. *It holds that*

$$\begin{aligned}\mathbb{E}(B_1^2|\boldsymbol{I} = (I_1, I_2)) &= 4a^2 I_1^2 + (1 + 4a + 4a^2)I_2^2 + 4a(1 + 3a + 2a^2)(I_1 + I_2)\\ &\quad + 4a(1 + 2a)I_1 I_2\,,\\ \mathbb{E}(B_2^2|\boldsymbol{I} = (I_1, I_2)) &= a^2(I_1 + I_2)^2 + a(1 + 3a + 2a^2)(I_1 + I_2)\,,\end{aligned}$$

where $a = \theta\rho/(\lambda_1 - \theta\rho)$.

Proof. Define $g_i = \mathbb{E}(B_i^2|\boldsymbol{I} = (1,0))$ and $h_i = \mathbb{E}(B_i^2|\boldsymbol{I} = (0,1))$ for $i = 1, 2$. Let a, p, m_i and k_i be given as in the proof of Lemma 1. Conditioning on the first event as before, we have $g_1 = p(1+2m_1+2k_1+2m_1k_1+g_1+h_1)$ and $h_1 = 1+2m_1+g_1$. Thus using the previous results, $g_1 = p(4 + 16a + 8a^2 + 2g_1)$ and $h_1 = 1 + 4a + g_1$. Since $p/(1 - 2p) = a$, it follows that $g_1 = 4a(1 + 4a + 2a^2)$ and $h_1 = 1 + 8a(1 + 2a + a^2)$. Similarly, we have $g_2 = p(1 + 4a + 2a^2 + 2g_2)$ and $h_2 = g_2$, which results in $g_2 = h_2 = a(1 + 4a + 2a^2)$. These imply that $\mathrm{Var}(B_1|\boldsymbol{I} = (1,0)) = \mathrm{Var}(B_1|\boldsymbol{I} = (0,1)) = g_1 - 4a^2 = 4a(1+3a+2a^2)$ and $\mathrm{Var}(B_2|\boldsymbol{I} = (1,0)) = \mathrm{Var}(B_2|\boldsymbol{I} = (0,1)) = g_2 - a^2 = a(1+3a+2a^2)$. Since individuals reproduce independently of each other, $\mathrm{Var}(B_1|\boldsymbol{I} = (I_1, I_2)) = 4a(1+3a+2a^2)(I_1+I_2)$ and $\mathrm{Var}(B_2|\boldsymbol{I} = (I_1, I_2)) = a(1+3a+2a^2)(I_1+I_2)$, which implies the lemma. □

Based on the construction of the processes described above, Theorem 4.1 and equation (4.3) in Ball & Donnelly (1995) yields that the probability, given B_1 and

B_2, that $\boldsymbol{Z_I}$ and $\boldsymbol{E_I^M}$ do not coincide is

$$\begin{aligned} p_{\boldsymbol{I},\boldsymbol{M}}^{(B_1,B_2)} &= 1 - \prod_{k=1}^{B_1}\left[1 - \frac{k-1}{M_1}\right]\prod_{l=1}^{B_2}\left[1 - \frac{l-1}{M_2}\right] \\ &\le 1 - \exp\left(-\frac{B_1(B_1-1)}{2M_1} - \frac{B_2(B_2-1)}{2M_2}\right) \\ &< \left(\frac{B_1(B_1-1)}{2M_1} + \frac{B_2(B_2-1)}{2M_2}\right), \end{aligned}$$

since $x > 1 - \exp(-x)$ for $x > 0$. Thus the corresponding unconditional probability $p_{\boldsymbol{I},\boldsymbol{M}}$ satisfies

$$p_{\boldsymbol{I},\boldsymbol{M}} \le \mathbb{E}\left(\frac{B_1(B_1-1)}{2M_1} + \frac{B_2(B_2-1)}{2M_2}\right),$$

so that Lemmas A.3 and A.4 imply that

$$p_{\boldsymbol{I},\boldsymbol{M}} = O(\max\{I_1, I_2\}^2 / \min\{M_1, M_2\}).$$

Hence the following result follows immediately.

Theorem A.2. *If* $\max\{I_1, I_2\}^2 / \min\{M_1, M_2\} \to 0$ *as* $\min\{M_1, M_2\} \to \infty$, *it follows that*

$$\lim_{\min\{M_1,M_2\}\to\infty} \mathbb{P}(\boldsymbol{E_I^M} = \boldsymbol{Z_I} \text{ for all } t \ge 0) = 1.$$

The result of the fourth paper can now be used to approximate the time to extinction for $\boldsymbol{Z_I}$ and thus for the epidemic process $\boldsymbol{E_I^M}$ based on coincidence of the processes. The application of the results of the following paper to the present case is illustrated after that paper.

Extinction times in multitype Markov branching processes

Dominik Heinzmann, *University of Zurich*

Abstract

In this paper, a distributional approximation to the time to extinction in a sub-critical continuous-time Markov branching process is derived. A limit theorem for this distribution is established and the error in the approximation is quantified. The accuracy of the approximation is illustrated in an epidemiological example. Since Markov branching processes serve as approximations to nonlinear epidemic processes in the initial and final stages, our results can also be used to describe the time to extinction for such processes.

Keywords: Multitype branching process, extinction time, convergence rate.

1. Introduction

This paper is concerned with approximating the time to extinction in a sub-critical multitype Markov branching process, starting with many individuals. The argument is based on the classical exponential approximation to the extinction probabilities (Athreya & Ney 1972, Harris 1963, Jagers 1975, Jagers et al. 2007, Sewastjanow 1974). These approximations are then combined with the branching property to derive a Gumbel approximation. The bound on the error in total variation distance is inversely proportional to a positive power of a weighted sum of the number of individuals of the different types. The power depends on the means and higher moments of the offspring distribution.

In infectious disease modeling, the initial and final stages of epidemic processes can often be approximated by suitable branching processes, see Whittle (1955). More recently Ball (1983), Ball & Donnelly (1995), Barbour & Utev (2004) and Barbour (2007) have used different constructions to quantify the path accuracy of such approximations. These results can be combined with ours to derive corresponding statements about the extinction time in epidemic processes.

2. Equations for extinction probabilities

The notation is chosen with Athreya & Ney (1972, p.200), Harris (1963, p.113) and Sewastjanow (1974, p.77) as basic references. For $k < \infty$, set $\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_k(t))$, where $Z_i(t)$ is the number of individuals of type i at time t. A type i individual has exponential lifetime with parameter a_i and rises at death j_i type i individuals, $1 \le i \le k$, with probability $p_i^{\boldsymbol{j}}$, for $\boldsymbol{j} = (j_1, \ldots, j_k) \in \mathbb{Z}_+^k$, independent of everything that has happened up to this time. Assume that

$$R_{il} := \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} j_l < \infty \quad \text{for } 1 \le i, l \le k\,. \tag{1}$$

Let $\|.\|$ be the supremum norm and let $\mathbb{P}_I$ be a conditional distribution of the process at time t given $\boldsymbol{Z}(0) = \boldsymbol{I}$, for $\boldsymbol{I} = (I_1, \ldots, I_k) \in \mathbb{Z}_+^k$. In particular, let $\mathbb{P}_i$ corresponds to the case when $Z_i(0) = 1$ and $Z_m(0) = 0$, $m \neq i$. Let T be the extinction time of the process and define the survival probability of the process when starting with a single type i individual as $q_i(t) := 1 - \mathbb{P}_i(T \leq t) = 1 - \mathbb{P}_i(\boldsymbol{Z}(t) = 0)$.

Then Harris (1963, p.114, equation 15.2) implies that

$$\frac{1}{a_i}\frac{d}{dt}(1 - q_i(t)) = q_i(t) - \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \left\{ 1 - \prod\nolimits_{l=1}^{k} [1 - q_l(t)]^{j_l} \right\} \quad \text{for } 1 \leq i \leq k\,. \tag{2}$$

Now the relation $(1 - x)^l \geq 1 - xl$ for $0 \leq x \leq 1$ together with the Bonferroni inequalities (Galambos & Simonelli 1996, p.27) imply that

$$\sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \left\{ 1 - \prod\nolimits_{l=1}^{k} [1 - q_l(t)]^{j_l} \right\} \leq \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \{\boldsymbol{j}^T \boldsymbol{q}(t)\} \tag{3}$$

and

$$\sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \left\{ 1 - \prod\nolimits_{l=1}^{k} [1 - q_l(t)]^{j_l} \right\} \geq \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \max\{\boldsymbol{j}^T \boldsymbol{q}(t) - F^{\boldsymbol{j}}(\boldsymbol{q}(t)), 0\}\,, \tag{4}$$

where $\boldsymbol{q}(t) = (q_1(t), \ldots, q_k(t))$, and

$$F^{\boldsymbol{j}}(\boldsymbol{q}(t)) := \frac{1}{2} \sum_{\substack{l,l'=1;\\ l \neq l'}}^{k} j_l j_{l'} q_l(t) q_{l'}(t) + \frac{1}{2} \sum_{l=1}^{k} j_l (j_l - 1) q_l^2(t) \leq \frac{1}{2} (\boldsymbol{j}^T \boldsymbol{q}(t))^2\,.$$

Using (3) and (4) in (2) and recalling (1) gives

$$\frac{1}{a_i}\frac{dq_i(t)}{dt} \leq \{(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{q}(t)\}_i \tag{5}$$

and

$$\frac{1}{a_i}\frac{dq_i(t)}{dt} = \{(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{q}(t)\}_i - v_i(t)\,, \tag{6}$$

where $v_i(t)$ summarizes all terms nonlinear in $\boldsymbol{q}(t)$ (see Section 5 for an example) and satisfies

$$\begin{aligned} 0 \leq v_i(t) &= \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \left\{ \sum_{l=1}^{k} [j_l q_l(t)] - 1 + \prod_{l=1}^{k} [1 - q_l(t)]^{j_l} \right\} \\ &\leq \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \min\{F^{\boldsymbol{j}}(\boldsymbol{q}(t)), \boldsymbol{j}^T \boldsymbol{q}(t)\}\,. \end{aligned} \tag{7}$$

Since equation (2) is nonlinear in $\boldsymbol{q}(t)$, it can in general not be solved analytically. However, we shall see using Theorem 3.1 that the behavior of the solution $\boldsymbol{q}(t)$ can be approximated by that of

$$\frac{d\boldsymbol{q}(t)}{dt} = \{\boldsymbol{A}(\boldsymbol{R}-\boldsymbol{I})\}\boldsymbol{q}(t) =: \boldsymbol{B}\boldsymbol{q}(t) \tag{8}$$

so long as $\|\boldsymbol{q}(t)\|$ is small and $\boldsymbol{A} := diag\{a_1, \ldots, a_k\}$. The matrix $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{R}-\boldsymbol{I})$ has non-negative elements off the diagonal, and is thus a Metzler-Leontief (ML) matrix (Seneta 1973, p.40). If $\boldsymbol{B}$ is irreducible (Seneta 1973, p.15), the process $\boldsymbol{Z}(t)$ is irreducible (Sewastjanow 1974, p.99), and the following Perron-Fröbenius result (Seneta 1973, Theorem 2.5) applies.

Theorem 2.1. *Assume that $\boldsymbol{B}$ is a $k \times k$ irreducible matrix with non-negative off-diagonal elements. Then there exists an eigenvalue ω such that:*

(i) *ω is real;*

(ii) *there are unique (up to a constant factor) strictly positive left $\boldsymbol{f}_1$ and right $\boldsymbol{b}_1$ eigenvectors associated with ω;*

(iii) *$\omega > Re(\omega_i)$ for any eigenvalue $\omega_i \neq \omega$ of $\boldsymbol{B}$;*

(iv) *ω is a simple root of the characteristic equation of $\boldsymbol{B}$.*

In what follows, it is assumed that the process is sub-critical, i.e. $\omega < 0$. Define $r := -\omega$. The left eigenvector $\boldsymbol{f}_1$ can be used to derive an upper bound for $\boldsymbol{q}(t)$ $(t > 0)$.

Lemma 2.2. *Assume that $\boldsymbol{f}_1^T = (f_{11}, \ldots, f_{1k})$ is such that $\|\boldsymbol{f}_1\| = 1$. Then*

$$q_i(t) \leq e^{-rt}\left[\frac{\boldsymbol{f}_1^T\mathbf{1}}{f_{1i}}\right], \quad \textit{for } 1 \leq i \leq k\,,$$

where $\mathbf{1}$ denotes a column vector of 1's.

Proof. Theorem 2.1 implies that $\boldsymbol{f}_1$ has only positive entries and hence inequality (5) implies that

$$\frac{d}{dt}\{\boldsymbol{f}_1^T\boldsymbol{q}(t)\} \leq \boldsymbol{f}_1^T\boldsymbol{B}\boldsymbol{q}(t) = \omega\boldsymbol{f}_1^T\boldsymbol{q}(t) = -r\boldsymbol{f}_1^T\boldsymbol{q}(t)\,,$$

and using Grönwall's lemma (Grönwall 1918/1919) yields

$$\boldsymbol{f}_1^T\boldsymbol{q}(t) \leq e^{-rt}\boldsymbol{f}_1^T\boldsymbol{q}(0) = e^{-rt}\boldsymbol{f}_1^T\mathbf{1}\,.$$

The result follows immediately, since $\boldsymbol{f}_1$ and $\boldsymbol{q}(t)$ are both positive vectors. □

The following useful lemma is proved by a standard argument.

Lemma 2.3. *Let X be a non-negative random variable with $\mathbb{E}(X) < \infty$ and let $d > 1$. Then for $\delta \to 0$, $\mathbb{E}(X\delta \wedge (X\delta)^d) = o(\delta)$. If in addition $\mathbb{E}(X^\psi) < \infty$ for some $1 \le \psi \le d$, then $\mathbb{E}(X\delta \wedge (X\delta)^d) \le 2\mathbb{E}(X\delta)^\psi = O(\delta^\psi)$.*

Let J_i denote a random variable with $\mathbb{P}(J_i = \boldsymbol{j}) = p_i^{\boldsymbol{j}}$.

Theorem 2.4. *If $\mathbb{E}(\|J_i\|) < \infty$ for $1 \le i \le k$, then $v_i(t) = o(\|\boldsymbol{q}(t)\|)$ as $t \to \infty$.*

Proof. From (7) it follows that

$$0 \le v_i(t) \le \sum_{\boldsymbol{j} \in \mathbb{N}^k} p_i^{\boldsymbol{j}} \left\{ \frac{1}{2} (\boldsymbol{j}^T \boldsymbol{q}(t))^2 \wedge (\boldsymbol{j}^T \boldsymbol{q}(t)) \right\} = \mathbb{E} \left\{ \frac{1}{2} (\boldsymbol{J}^T \boldsymbol{q}(t))^2 \wedge (\boldsymbol{J}^T \boldsymbol{q}(t)) \right\}, \quad (9)$$

where $\boldsymbol{J}^T = (J_1, \ldots, J_k)$. Since $\|\boldsymbol{q}(t)\| \le \sum_{i=1}^k |q_i(t)|$, Lemma 2.2 indicates that $\|\boldsymbol{q}(t)\| \to 0$ as $t \to \infty$, and thus Lemma 2.3 can be applied. □

The following corollary gives a more specific asymptotic upper bound on $v_i(t)$, if the offspring distributions have a finite moment higher than the first.

Corollary 2.5. *Suppose that $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ for some $0 < \alpha \le 1$ and for all $1 \le i \le k$. Then there exist constants $c_i^* < \infty$ such that $0 \le v_i(t) \le c_i^* \|\boldsymbol{q}(t)\|^{1+\alpha}$, $1 \le i \le k$.*

Proof. The proof follows immediately from (9) and from Lemma 2.3. □

3. Exponential limit behavior

The following result is the basis for approximating the survival time of the process, bounding the error in the exponential approximation to the extinction probabilities

Theorem 3.1. *Assume that $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ for some $0 < \alpha \le 1$ and all $1 \le i \le k$. If $\boldsymbol{B}$ is irreducible with largest eigenvalue $-r < 0$, the probability of survival $q_i(t)$ when starting with a single individual of type i satisfies*

$$q_i(t) = c_i e^{-rt}(1 + o(e^{-\gamma t})),$$

where $0 < \gamma < r$ is given below, and $c_i/c_l = b_{1i}/b_{1l}$, where $\boldsymbol{b}_1$ is the right eigenvector of $\boldsymbol{B}$ corresponding to the eigenvalue $-r$.

Proof. Let $\boldsymbol{v}(t) := (v_1(t), \ldots, v_k(t))$ and define $\boldsymbol{u}(t) := e^{rt}\boldsymbol{q}(t)$. It follows from (6) that

$$\frac{d}{dt}\boldsymbol{u}(t) = \boldsymbol{C}\boldsymbol{u}(t) - e^{rt}\boldsymbol{A}\boldsymbol{v}(t), \quad (10)$$

where the largest eigenvalue of $\boldsymbol{C} := (\boldsymbol{B} + r\boldsymbol{I})$ is 0. Let 0 and $\{\omega_j;\, 2 \le j \le k^*\}$ denote the eigenvalues corresponding to the $k^* \le k$ Jordan blocks of $\boldsymbol{C}$, and denote

by k_j, $2 \le j \le k^*$, their dimensions. The left eigenvector of $\boldsymbol{C}$ corresponding to the eigenvalue 0 is $\boldsymbol{f}_1^T$; for $1 \le m \le k_j$, and $2 \le j \le k^*$, let $\boldsymbol{f}_{i,m}^T$ denote the corresponding Jordan basis vectors, with $\|\boldsymbol{f}_{j,m}\| = 1$; set $-\beta_j = Re(\omega_j)$, so that for $2 \le m \le k_j$ and $2 \le j \le k^*$, $\boldsymbol{f}_{j,1}^T\boldsymbol{C} = \omega_j \boldsymbol{f}_{j,1}^T$ and $\boldsymbol{f}_{j,m}^T\boldsymbol{C} = \omega_j \boldsymbol{f}_{j,m}^T + \boldsymbol{f}_{j,m-1}^T$

Define $\boldsymbol{w}(t) := (\boldsymbol{f}_1^T\boldsymbol{A}\boldsymbol{v}(t))/(\boldsymbol{f}_1^T\boldsymbol{q}(t))$. From Lemma 2.2 and Corollary 2.5, it is immediate that $\|\boldsymbol{w}(t)\| = O(e^{-r\alpha t})$ and hence that $\int_s^\infty \|\boldsymbol{w}(t)\|dt < \infty$.

Now, from (10),

$$\frac{d}{dt}\log(\boldsymbol{f}_1^T\boldsymbol{u}(t)) = -\boldsymbol{w}(t)\,,$$

and hence

$$\log(\boldsymbol{f}_1^T\boldsymbol{q}(t)) + rt = \log(\boldsymbol{f}_1^T\boldsymbol{q}(s)) + rs - \int_s^t \boldsymbol{w}(z)dz\,.$$

By the Cauchy criterion and the integrability of $\|\boldsymbol{w}(z)\|$ it follows that

$$\lim_{t\to\infty}\{\log(\boldsymbol{f}_1^T\boldsymbol{q}(t)) + rt\} =: \log h^*$$

exists and is finite and thus, using $\|\boldsymbol{w}(t)\| = O(e^{-r\alpha t})$,

$$\boldsymbol{f}_1^T\boldsymbol{q}(t) = h^* e^{-rt}(1 + O(e^{-r\alpha t}))$$

with $h^* > 0$.

For the remaining part of the argument, we refer to the theory of perturbed linear systems. Rewrite (10) as

$$\frac{d}{dt}\boldsymbol{u}(t) = [\boldsymbol{C} + \boldsymbol{D}(t)]\boldsymbol{u}(t)\,, \tag{11}$$

where

$$\boldsymbol{D}(t) = -\frac{\boldsymbol{A}\boldsymbol{v}(t)\boldsymbol{q}(t)^T}{\|\boldsymbol{q}(t)\|^2}\,,$$

so that $\|\boldsymbol{D}(t)\| \le K^* e^{-r\alpha t}$, with $K^* < \infty$. System (11) is a special case of the system in Theorem 2 of Levinson (1948), from which it follows that, for any $\gamma < \min\{r\alpha, \beta_{[2]}\}$, where $-\beta_{[2]}$ is the second largest real part of any eigenvalue of $\boldsymbol{C}$, we have $|\boldsymbol{f}_{j,m}^T\boldsymbol{u}(t)| = o(e^{-\gamma t})$, $1 \le m \le k_j$, $2 \le j \le k^*$.

Now the set of vectors $\{\boldsymbol{f}_1^T, \boldsymbol{f}_{j,m}^T;\ 1 \le m \le k_j,\ 2 \le j \le k^*\}$ constitutes a basis of $\mathbb{R}^d$. Let $\boldsymbol{x} \in \mathbb{R}^d$ have coefficients $(x_1, x_{j,m};\ 1 \le m \le k_j, 2 \le j \le k^*)$ with respect to this basis. Then

$$\boldsymbol{x}^T(e^{rt}\boldsymbol{q}(t)) = (x_1\boldsymbol{f}_1^T + \sum_{j=2}^{k^*}\sum_{n=1}^{k_j} x_{j,n}\boldsymbol{f}_{j,n}^T)\boldsymbol{u}(t) = x_1 h^* + o(\|\boldsymbol{x}\|e^{-\gamma t})\,. \tag{12}$$

In particular, for $1 \le i \le k$, it follows that $q_i(t) = c_i e^{-rt}(1 + o(e^{-\gamma t}))$, where

$$c_i = (\boldsymbol{e}_i^T\boldsymbol{b}_1)h^* = b_{1i}h^* > 0\,, \tag{13}$$

with $\boldsymbol{e}_i$ a column vector with a 1 at the position i and 0's else and $\boldsymbol{b}_1$ the right eigenvector of $\boldsymbol{B}$ corresponding to the eigenvalue $-r$ such that $\boldsymbol{f}_1^T\boldsymbol{b}_1 = 1$. □

Remark 3.2. *The order of convergence is simplified for clarity in the statement of Theorem 3.1. For the case where $\boldsymbol{B}$ is diagonalizable, the exact formulation is as follows. If $-\beta_2$ is the second largest real part of an eigenvalue of $\boldsymbol{C}$ and if $r\alpha \neq \beta_2$, then $q_i(t) = c_i e^{-rt}(1 + O(e^{-\gamma t}))$ where $\gamma = \min\{r\alpha, \beta_2\}$. Otherwise if $r\alpha = \beta_2$, then $q_i(t) = c_i e^{-rt}(1 + O(te^{-r\alpha t}))$.*

4. Time to extinction

If $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ for some $0 < \alpha \leq 1$ and all $1 \leq i \leq k$, Theorem 3.1 implies that, as $t \to \infty$,

$$\mathbb{P}_{\boldsymbol{I}}(T > t) = 1 - \prod_{i=1}^{k}[\mathbb{P}_i(T \leq t)]^{I_i} = 1 - \prod_{i=1}^{k}[1 - q_i(t)]^{I_i} \sim 1 - \prod_{i=1}^{k}[1 - c_i e^{-rt}]^{I_i}\,, \quad (14)$$

where $c_i > 0$ $(1 \leq i \leq k)$, $\boldsymbol{I} = (I_1, \ldots, I_k)$ with I_i the initial number of type i individuals. Define $C_{\boldsymbol{I}} := \sum_{j=1}^{k} I_j c_j$. The approximation error in (14) is controlled by the following result.

Lemma 4.1. *Suppose that $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ for some $0 < \alpha \leq 1$ and all $1 \leq i \leq k$. Then, for any γ as in Theorem 3.1, there exist $t_0, \nu_1 < \infty$, not depending on $\boldsymbol{I}$, such that*

$$\left|\prod_{i=1}^{k}[1 - q_i(t)]^{I_i} - \prod_{i=1}^{k}[1 - c_i e^{-rt}]^{I_i}\right| \leq \nu_1 C_{\boldsymbol{I}} e^{-\frac{1}{2}C_{\boldsymbol{I}}e^{-rt}} e^{-(r+\gamma)t}\,, \quad t \geq t_0\,.$$

Proof. Denote the approximation error in (14) as $\epsilon^{(1)}(t)$. Choose

$$t_1 \geq (1/r)\max_{1\leq i\leq k}(\log c_i)_+$$

such that $q_i(t) \geq (1/2)c_i e^{-rt}$ for all i and $t \geq t_1$. Using

$$\left|\prod_{i=1}^{k} A_i - \prod_{i=1}^{k} B_i\right| \leq \sum_{l=1}^{k} |A_l - B_l| \left(\prod_{i=1}^{l-1} |A_i|\right)\left(\prod_{i=l+1}^{k} |B_i|\right)\,,$$

with $A_i = [1 - q_i(t)]^{I_i}$ and $B_i = [1 - c_i e^{-rt}]^{I_i}$, it follows that

$$\epsilon^{(1)}(t) \leq e^{-\frac{1}{2}C_{\boldsymbol{I}}e^{-rt}} \sum_{i=1}^{k} I_i \left\{\frac{|[1 - q_i(t)] - [1 - c_i e^{-rt}]|}{\min(1 - q_i(t), 1 - c_i e^{-rt})}\right\}\,, \quad t \geq t_1\,.$$

Determine t_2 such that $\min_{1\le i\le k}\{\min(1-q_i(t), 1-c_ie^{-rt})\} \ge 1/2$ for $t \ge t_2$. From Theorem 3.1, we have $|q_i(t) - c_ie^{-rt}| \le K^*c_ie^{-(r+\gamma)t}$, $1 \le i \le k$, for some $K^* < \infty$. Hence for all $t \ge t_0 := \max\{t_1, t_2\}$,

$$\epsilon^{(1)}(t) \le 2C_{\boldsymbol{I}}e^{-\frac{1}{2}C_{\boldsymbol{I}}e^{-rt}}K^*e^{-(r+\gamma)t}\,,$$

for γ as in Theorem 3.1. □

A further approximation to the last term in (14) is

$$1 - \prod_{i=1}^{k}[1 - c_ie^{-rt}]^{I_i} \sim 1 - e^{-C_{\boldsymbol{I}}e^{-rt}}\,. \tag{15}$$

The approximation error in (15) can be bounded as follows (the proof is omitted).

Lemma 4.2. *We have*

$$\left|\prod_{i=1}^{k}[1 - c_ie^{-rt}]^{I_i} - e^{-C_{\boldsymbol{I}}e^{-rt}}\right| \le \nu_2 C_{\boldsymbol{I}}e^{-C_{\boldsymbol{I}}e^{-rt}}e^{-2rt}\ ,\ \ t \ge t_2\,,$$

where t_2 is as for Lemma 4.1 and $\nu_2 = \max_{1\le i\le k} c_i < \infty$.

Remark 4.3. *Lemmas 4.1 and 4.2 imply that*

$$\epsilon^{(1)}(t) \le \frac{\nu_1}{C_{\boldsymbol{I}}^{\gamma/r}}\max_{x>0}\left\{e^{-x/2}x^{1+\gamma/r}\right\} = \frac{\nu_3}{C_{\boldsymbol{I}}^{\gamma/r}}\ ,\ \ t \ge t_0\,;$$
$$\epsilon^{(2)}(t) \le \frac{4\nu_2e^{-2}}{C_{\boldsymbol{I}}} = \frac{\nu_4}{C_{\boldsymbol{I}}}\ ,\ \ t \ge t_2\,,$$

with t_0, t_2, ν_1, ν_2 and γ as before.

Definition 4.4. *Define the random variable $\tilde{T}_{\boldsymbol{I}}$ such that $\mathbb{P}(\tilde{T}_{\boldsymbol{I}} > t) = 1 - \exp(-C_{\boldsymbol{I}}\exp(-rt))$, where $C_{\boldsymbol{I}} = \sum_{i=1}^{k} I_ic_i$. The random variable $\tilde{T}_{\boldsymbol{I}}$ satisfies*

$$\tilde{T}_{\boldsymbol{I}} = \frac{\log C_{\boldsymbol{I}}}{r} + \frac{1}{r}V\,,$$

where V has a Gumbel distribution.

Theorem 4.5. *Suppose $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ for some $0 < \alpha \le 1$ and for all $1 \le i \le k$. Then for $t \ge 0$ and with $\gamma < r\alpha$ as in Theorem 3.1, there is a constant $\nu^* < \infty$ such that*

$$|\mathbb{P}_{\boldsymbol{I}}(T > t) - \mathbb{P}(\tilde{T}_{\boldsymbol{I}} > t)| \le \frac{\nu^*}{C_{\boldsymbol{I}}^{\gamma/r}}\,.$$

Proof. Remark 4.3 implies that

$$|\mathbb{P}_{\boldsymbol{I}}(T > t) - \mathbb{P}(\tilde{T}_{\boldsymbol{I}} > t)| \leq \frac{\nu_3}{C_{\boldsymbol{I}}^{\gamma/r}} + \frac{\nu_4}{C_{\boldsymbol{I}}}\ ,\ t \geq t_0\,.$$

For $t \leq t_0$, we have

$$0 \leq \mathbb{P}_{\boldsymbol{I}}(T \leq t) \leq \mathbb{P}_{\boldsymbol{I}}(T \leq t_0) \leq \mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t_0) + \frac{\nu_3}{C_{\boldsymbol{I}}^{\gamma/r}} + \frac{\nu_4}{C_{\boldsymbol{I}}}\,,$$

and

$$0 \leq \mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t) \leq \mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t_0) = e^{-C_{\boldsymbol{I}} e^{-rt_0}}\,,$$

completing the proof. □

Theorem 4.5 thus shows that

$$d_K\left(\mathscr{L}\left(T - \frac{\log C_{\boldsymbol{I}}}{r} \mid Z(0) = I\right), \mathscr{L}\left(\frac{V}{r}\right)\right) = O(C_{\boldsymbol{I}}^{-\gamma/r})$$

as $\|I\| \to \infty$, where d_K denotes the Kolmogorov distance between the distributions indicated by $\mathscr{L}$, and γ is as in Theorem 3.1.

We now strengthen the mode of convergence. Let $\tilde{f}_{\boldsymbol{I}}$ the probability density function of $\tilde{T}_{\boldsymbol{I}}$, and let $f_{\boldsymbol{I}}$ that of T under $\mathbb{P}_{\boldsymbol{I}}$.

Lemma 4.6. *Suppose* $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ *for some* $0 < \alpha \leq 1$ *and all* $1 \leq i \leq k$. *For all* $t \geq t_0$, *there exists a constant* $K < \infty$ *such that*

$$|f_{\boldsymbol{I}}(t) - \tilde{f}_{\boldsymbol{I}}(t)| \leq K C_{\boldsymbol{I}} e^{-(r+\gamma)t}(1 + C_{\boldsymbol{I}} e^{-rt}) e^{-\frac{1}{2} C_{\boldsymbol{I}} e^{-rt}}\,,$$

where γ *is as in Theorem 3.1 and* t_0 *as for Lemma 4.1.*

Proof. From (14) we know that $\mathbb{P}_{\boldsymbol{I}}(T \leq t) = \prod_{i=1}^k [1 - q_i(t)]^{I_i}$, and thus

$$f_{\boldsymbol{I}}(t) = \frac{d}{dt}\mathbb{P}_{\boldsymbol{I}}(T \leq t) = \mathbb{P}_{\boldsymbol{I}}(T \leq t) \sum_{i=1}^k \left[-\frac{I_i}{1 - q_i(t)} \frac{dq_i(t)}{dt}\right]. \tag{16}$$

Furthermore,

$$\tilde{f}_{\boldsymbol{I}}(t) = \frac{d}{dt}\mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t) = \mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t) r C_{\boldsymbol{I}} e^{-rt}\,. \tag{17}$$

Lemmas 4.1 and 4.2 imply that, for $t \geq t_0$,

$$|\mathbb{P}_{\boldsymbol{I}}(T \leq t) - \mathbb{P}(\tilde{T}_{\boldsymbol{I}} \leq t)| \leq K_1 C_{\boldsymbol{I}} e^{-\frac{1}{2} C_{\boldsymbol{I}} e^{-rt}} e^{-(r+\gamma)t}\,, \tag{18}$$

for some $K_1 < \infty$. Then, also for $t \geq t_0$,

$$\begin{aligned}\left|\boldsymbol{I}^T \frac{d\boldsymbol{q}(t)}{dt} - \sum_{i=1}^k \frac{I_i}{1 - q_i(t)} \frac{dq_i(t)}{dt}\right| &\leq \sum_{i=1}^k I_i \left|\frac{dq_i(t)}{dt}\right| \left|\frac{1}{1 - q_i(t)} - 1\right| \\ &\leq K_2 C_{\boldsymbol{I}} e^{-2rt}\,,\end{aligned} \tag{19}$$

with $K_2 < \infty$, since Lemma 2.2 and Corollary 2.5 imply that $|\boldsymbol{I}^T d\boldsymbol{q}(t)/dt| \leq K_3 C_{\boldsymbol{I}} e^{-rt}$, with $K_3 < \infty$, and for $t \geq t_0$, $1 - q_i(t) \geq 1/2$ for $1 \leq i \leq k$, implying that

$$\left| \frac{1}{1 - q_i(t)} - 1 \right| \leq 2q_i(t) = O(e^{-rt}) .$$

Now, since $d\boldsymbol{q}(t)/dt = \boldsymbol{B}q(t) - \boldsymbol{A}\boldsymbol{v}(t)$ as in (6), we have

$$|\boldsymbol{I}^T \frac{d\boldsymbol{q}(t)}{dt} + rC_{\boldsymbol{I}} e^{-rt}| = |\boldsymbol{I}^T \boldsymbol{B}\boldsymbol{q}(t) - \boldsymbol{I}^T \boldsymbol{A}\boldsymbol{v}(t) + rC_{\boldsymbol{I}} e^{-rt}| \leq K_4 C_{\boldsymbol{I}} e^{-(r+\gamma)t} . \qquad (20)$$

The final inequality in (20) with $K_4 < \infty$ follows because:

a) Equation (12) implies that

$$|\boldsymbol{I}^T \boldsymbol{B}\boldsymbol{q}(t) - \boldsymbol{I}^T \boldsymbol{B}\boldsymbol{b}_1 h^* e^{-rt}| \leq K_5 C_{\boldsymbol{I}} e^{-(r+\gamma)t} \quad \text{for } t \geq 0 \text{ and } K_5 < \infty ;$$

b) Equation (13) and the definition of $\boldsymbol{b}_1$ give $\boldsymbol{I}^T \boldsymbol{B}\boldsymbol{b}_1 h^* = -C_{\boldsymbol{I}} r$; and

c) Corollary 2.5 shows that

$$|\boldsymbol{I}^T \boldsymbol{A}\boldsymbol{v}(t)| \leq K_6 C_{\boldsymbol{I}} e^{-r(1+\alpha)t} \quad \text{for } t \geq 0 \text{ and } K_6 < \infty .$$

Combining (19) and (20) thus gives

$$\left| rC_{\boldsymbol{I}} e^{-rt} + \sum_{i=1}^{k} \frac{I_i}{1 - q_i(t)} \frac{dq_i(t)}{dt} \right| \leq K_7 e^{-(r+\gamma)t} . \qquad (21)$$

Using (18) and (21), together with the triangle inequality now applied to the difference of (16) and (17) in the form

$$|A_1 A_2 - B_1 B_2| \leq |A_1 - B_1||A_2 - B_2| + |B_2||A_1 - B_1| + |B_1||A_2 - B_2|$$

yields the lemma. □

Using Lemma 4.6, we can show that the distribution of T under $\mathbb{P}_{\boldsymbol{I}}$ can be well approximated by that of $\tilde{T}_{\boldsymbol{I}}$ in terms of probability densities and also in the total variation distance d_{TV}.

Theorem 4.7. *Suppose* $\mathbb{E}(\|J_i\|^{1+\alpha}) < \infty$ *for some* $0 < \alpha \leq 1$ *and all* $1 \leq i \leq k$. *Then there exist constants* $K_a, K_b < \infty$ *such that*

(*i*) $|f_{\boldsymbol{I}}(t) - \tilde{f}_{\boldsymbol{I}}(t)| \leq K_a C_{\boldsymbol{I}}^{-\frac{\gamma}{r}}$, $t \geq 0$;

(*ii*) $d_{TV}(\mathscr{L}(T|\boldsymbol{Z}(0) = \boldsymbol{I}), \mathscr{L}(\tilde{T}_{\boldsymbol{I}})) = \frac{1}{2} \int_0^\infty |f_{\boldsymbol{I}}(t) - \tilde{f}_{\boldsymbol{I}}(t)| dt \leq K_b C_{\boldsymbol{I}}^{-\frac{\gamma}{r}}$.

Proof. For $t \geq t_0$, (i) follows from Lemma 4.6, since $x^{1+\frac{\gamma}{r}}(1+x)e^{-\frac{x}{2}}$ is uniformly bounded in $x \geq 0$. For $t \leq t_0$, we have

$$\tilde{f}_{\boldsymbol{I}}(t) \leq C_{\boldsymbol{I}} r e^{-C_{\boldsymbol{I}} e^{-rt_0}} = O(C_{\boldsymbol{I}}^{-s}) \quad \text{for all } \; s > 0\,;$$

similarly, from (14) and (16), it can be shown that

$$f_{\boldsymbol{I}}(t) \leq K C_{\boldsymbol{I}} e^{-\sum_{i=1}^{k} d_i I_i}$$

with $d_i = -\log(1 - q_i(t_0)) > 0$ $(1 \leq i \leq k)$ and $K < \infty$, which is also of order $O(C_{\boldsymbol{I}}^{-s})$ for all s, completing the proof of part (i).

For part (ii), by Lemma 4.6,

$$\int_{t_0}^{\infty} |f_{\boldsymbol{I}}(t) - \tilde{f}_{\boldsymbol{I}}(t)| dt \leq K_c C_{\boldsymbol{I}}^{-\frac{\gamma}{r}}\,,$$

for $K_c < \infty$ a constant. The remaining part is bounded using part (i). □

5. Application

Theorem 4.5 is illustrated by a two-type model for parasitic resistance, in which the parasite can enter a resting phase during which it does not reproduce, but can be transmitted easily to a new host. An example is the transmission cycle of the parasitic protozoa *Toxoplasma gondii* (Eckert et al. 2005) in the intermediate hosts, which are warm-blooded. One third of the world's human population is estimated to carry a Toxoplasma infection (Montoya & Liesenfeld 2004). The growth rate of a parasite population within the intermediate host can be modelled by a two-type continuous-time Markov branching process. A parasite is of type 1 if it is in the active state, and of type 2 if it is in the resting state. A type 1 parasite can either die at rate d_1, enter the resting state at rate r_1 or reproduce itself by binary splitting at rate ρ. A type 2 parasite can either die at rate d_2 or becomes active within the host by changing to the active state at rate r_2. The transmission of the parasite to another host is incorporated in the death event. All inter-event times are exponentially distributed.

Let $Z_i = Z_i(t)$ $(i = 1, 2)$ be the number of type i parasites in a host at time $t \geq 0$. The transition scheme of the process is

Transition	**Rate**
$Z_1 \to Z_1 - 1,\ Z_2 \to Z_2$	$d_1 Z_1$
$Z_1 \to Z_1 - 1,\ Z_2 \to Z_2 + 1$	$r_1 Z_1$
$Z_1 \to Z_1 + 1,\ Z_2 \to Z_2$	ρZ_1
$Z_1 \to Z_1,\ \quad Z_2 \to Z_2 - 1$	$d_2 Z_2$
$Z_1 \to Z_1 + 1,\ Z_2 \to Z_2 - 1$	$r_2 Z_2$

(22)

Table 1: Accuracy of the extinction time approximation of the two-type Markov branching process (22). For given $\boldsymbol{I} = (I_1, I_2)$, the approximation $\tilde{T}_{\boldsymbol{I}}$ is compared to T ($500'000$ simulations) by computing the proportion of simulated values of T larger or equal to the median $\tilde{m}$ of $\tilde{T}_{\boldsymbol{I}}$, and by calculating the proportion of simulated values of T falling into the interquartile range (IQR) defined as the interval $(\tilde{q}_1, \tilde{q}_3)$, where $\tilde{q}_1$ and $\tilde{q}_3$ are the first and third quartiles of the approximating distribution of $\tilde{T}_{\boldsymbol{I}}$. The results are displayed for $I_1 : I_2 = 5 : 1$ (upper table) and $2 : 1$ (lower table) and different values of I_1. The corresponding $C_{\boldsymbol{I}} = c_1 I_1 + c_2 I_2$ are also represented. The model parameters are set to $(d_1, r_1, \rho) = (1, 1, 0.5)$ and $(d_2, r_2) = (1.2, 0.8)$ such that $-r < 0$.

$I_1 = 5I_2$	10	50	100	500	1000	5000
$C_{\boldsymbol{I}}$	9.618	48.090	96.179	480.897	961.793	4808.966
$\mathbb{P}(\geq \tilde{m})$	0.516	0.503	0.501	0.501	0.498	0.500
IQR	0.529	0.506	0.504	0.501	0.500	0.500

$I_1 = 2I_2$	10	50	100	500	1000	5000
$C_{\boldsymbol{I}}$	11.342	56.711	113.422	567.111	1134.221	5671.105
$\mathbb{P}(\geq \tilde{m})$	0.514	0.503	0.502	0.499	0.500	0.500
IQR	0.525	0.504	0.503	0.501	0.499	0.500

Let $a_1 := d_1 + r_1 + \rho$ and $a_2 := d_2 + r_2$ be the total rates of transition for type 1 and 2 individuals, respectively. For $q_i(t)$ $(i = 1, 2)$, system (2) yields

$$\frac{d\boldsymbol{q}(t)}{dt} = \begin{pmatrix} (-a_1 + 2\rho) & r_1 \\ r_2 & -a_2 \end{pmatrix} \boldsymbol{q}(t) - \begin{pmatrix} \rho q_1(t)^2 \\ 0 \end{pmatrix} = \boldsymbol{B}\boldsymbol{q}(t) - \boldsymbol{v}(t)\,. \tag{23}$$

Since $\|\boldsymbol{J}\| \leq 2$ for $\boldsymbol{Z}(0) = (1, 0)^T$ and $\boldsymbol{Z}(0) = (0, 1)^T$, we can take $\alpha = 1$.

Theorem 2.1 implies that $\boldsymbol{B}$ has a unique real largest eigenvalue $-r$, with corresponding positive left $\boldsymbol{f}_1^T$ and right $\boldsymbol{b}_1$ eigenvectors, which are given by $-r = (-(a_1 - 2\rho + a_2) + \sqrt{D})/2$ and

$$\boldsymbol{f}_1^T = \frac{1}{N_1}\left(\frac{a_2 + 2\rho - a_1 + \sqrt{D}}{2r_1}, 1\right) \;,\; \boldsymbol{b}_1^T = \frac{1}{N_2}\left(\frac{a_2 + 2\rho - a_1 + \sqrt{D}}{2r_2}, 1\right) ,$$

where $D = ((a_1 - 2\rho + a_2)^2 - 4(a_2(a_1 - 2\rho) - r_1 r_2))$, N_1 and N_2 are appropriate constants such that $|\boldsymbol{f}_1| = 1$ and $\boldsymbol{f}_1^T \boldsymbol{b}_1 = 1$.

The process $(\boldsymbol{Z}(t))_{t\geq 0}$ is sub-critical if and only if (i) $a_2(a_1 - 2\rho) > r_1 r_2$ and (ii) $a_1 - 2\rho + a_2 > 0$. Let the model parameter be fixed as in Table 1 such that the process is sub-critical. Thus $r = 0.821$, $\beta_2 = 1.857$, $\alpha = 1$ and Remark 3.2

implies that $q_i(t) = c_i e^{-rt}(1 + O(e^{-rt}))$. Furthermore, the Kolmogorov and the total variation distance between the distributions of T given $\boldsymbol{Z}(0) = \boldsymbol{I}$ and of $\tilde{T}_{\boldsymbol{I}}$ are both of order $O(C_{\boldsymbol{I}}^{-1})$. To compute $c_i = (\boldsymbol{e}_i^T \boldsymbol{b}_1) h^*$ $(i = 1, 2)$, it is necessary to determine h^*, given by

$$\log h^* := \lim_{t \to \infty} \{\log(\boldsymbol{f}_1^T \boldsymbol{q}(t)) + rt\} .$$

This entails numerical solution of the system (23) up to a sufficient large t. To increase numerical stability, it is advisable to solve for $e^{rt}\boldsymbol{q}(t)$ instead of $\boldsymbol{q}(t)$ by appropriately reformulating (23). To determine an appropriate t, the reformulated system is successively solved for $t \in \{10, 11, 12, \ldots\}$ and the corresponding c_1, c_2 are evaluated until the absolute differences of successive values of c_1 and c_2 are both smaller than some pre-defined level, 10^{-10} in our example, resulting in $c_1 = 0.847$ $c_2 = 0.575$ at $t = 16$.

Given $\boldsymbol{Z}(0) = \boldsymbol{I} = (I_1, I_2)$, the distribution of $\tilde{T}_{\boldsymbol{I}}$ given in Definition 4.4 can be compared with the distribution of the true extinction time T, which has to be computed by simulation, since the exact result is inaccessible. For the simulation, the Markov chain (22) can be simulated by the classical Gillespie algorithm (Gillespie 1977) or an improved version thereof (Gillespie & Petzold 2003).

Table 1 indicates a location and a scale measure for evaluating the approximation performance. The closeness of the probabilities $\mathbb{P}_{\boldsymbol{I}}(T > \tilde{m})$ and $\mathbb{P}_{\boldsymbol{I}}(\tilde{q}_1 < T < \tilde{q}_3)$ to their limiting values 0.5, where $\tilde{m}$, $\tilde{q}_1$ and $\tilde{q}_3$ are the median and the first and third quartiles of the approximating distribution of $\tilde{T}_{\boldsymbol{I}}$, increases with higher values of $C_{\boldsymbol{I}}$, which is in line with the previous results. Figure 1 represents the density function of the approximated extinction time versus the true one for different initial configurations $\boldsymbol{I} = (I_1, I_2)$. The density of the approximated distribution closely matches the distribution of the simulated times, supporting the results in this paper.

Acknowledgements

The author wish to thank Andrew Barbour for many helpful suggestions and comments. The author also gratefully acknowledge the referee for suggestions that greatly improved the presentation. This work was supported by the Schweizerischer Nationalfonds (SNF), project no. 107726.

References

Athreya, K. B. & Ney, P. E. (1972), *Branching Processes*, Berlin: Springer.

Ball, F. (1983), 'The threshold behaviour of epidemic models', *J. Appl. Prob.* **20**, 227–241.

Ball, F. & Donnelly, P. (1995), 'Strong approximations for epidemic models', *Stoch. Procs. Applics.* **55**, 1–25.

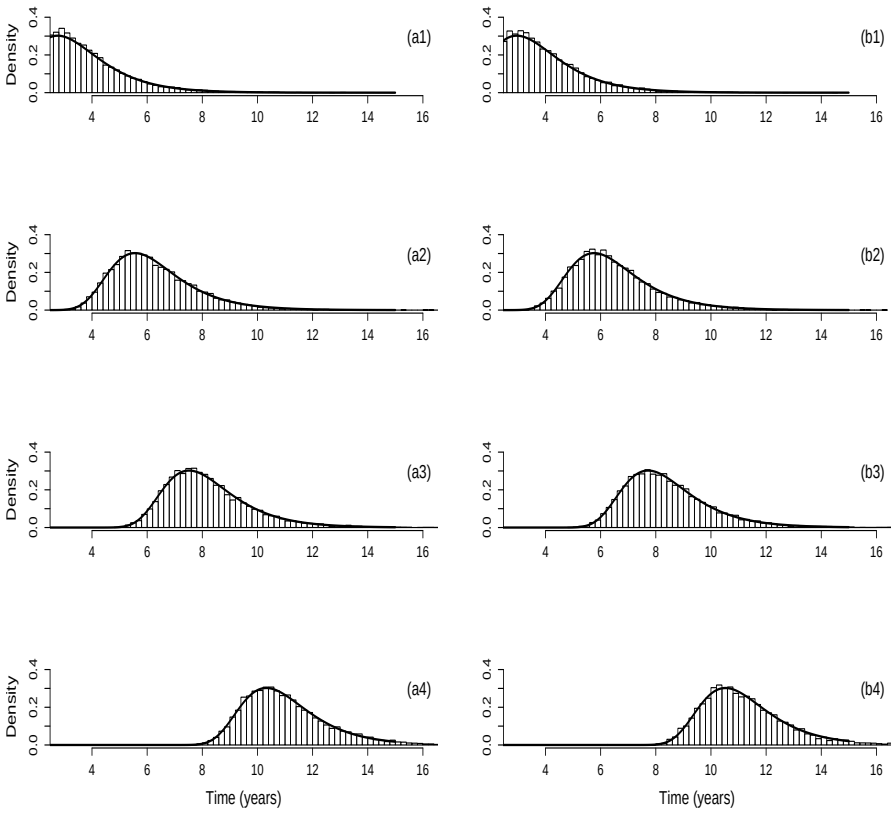

Figure 1: *Density distribution of* $\tilde{T}_{\boldsymbol{I}}$ *(solid line) versus the simulated distribution of* T *(histogram of* 10000 *simulations) for ratios* $I_1 : I_2 = 5 : 1$ ((a1)-(a4)) *and* $2 : 1$ ((b1)-(b4)) *with* I_1 *equal to* 10 ((a1) and (b1)), 100 ((a2) and (b2)), 500 ((a3) and (b3)) *and* 5000 ((a4) and (b4)). *The model parameters are fixed as in Table 1.*

Barbour, A. D. (2007), 'Coupling a branching process to an infinite dimensional epidemic process'.
URL: *arXiv:math.PR/0710.3697v1*

Barbour, A. D. & Utev, S. (2004), 'Approximating the Reed-Frost epidemic process', *Stoch. Procs. Applics.* **113**, 173–197.

Eckert, J., T., F. K., Zahner, H. & Deplazes, P. (2005), *Lehrbuch der Parasitologie für die Tiermedizin*, 1 edn, Stuttgart:Enke Verlag.

Galambos, J. & Simonelli, I. (1996), *Bonferroni-Type Inequalities with Applications*, New York: Springer.

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', *J Chem Phys* **81**, 2340–2361.

Gillespie, D. T. & Petzold, L. R. (2003), 'Improved leap-size selection for accelerated stochastic simulation', *J Chem Phys* **119**, 8229–8234.

Grönwall, T. H. (1918/1919), 'Note on the derivatives with respect to a parameter of the solutions of a system of differential equations', *Ann. Math.* **20**, 292–296.

Harris, T. E. (1963), *The theory of branching processes*, Berlin: Springer.

Jagers, P. (1975), *Branching Processes with Biological Applications*, New York: Wiley and Sons.

Jagers, P., Klebaner, F. C. & Sagitov, S. (2007), 'On the path to extinction', *PNAS* **104**, 6107–6111.

Levinson, N. (1948), 'The asymptotic nature of solutions of linear systems of differential equations', *Duke Math. J.* **15**, 111–126.

Montoya, J. G. & Liesenfeld, O. (2004), 'Toxoplasmosis', *The Lancet* **363**, 1965–1976.

Seneta, E. (1973), *Non-negative matrices: An introduction to theory and applications*, 1 edn, George Allen and Unwin Ltd.

Sewastjanow, B. A. (1974), *Verzweigungsprozesse*, Berlin: Akademie-Verlag.

Whittle, P. (1955), 'The outcome of a stochastic epidemic - A note on Bailey's paper', *Biometrika* **42**, 116–122.

Coupling of an epidemic model to a branching process: Application

The final part of the dissertation illustrates the application of the results in the preceding paper to the prevalence-based model introduced in the Chapter "Coupling of an epidemic model to a branching process: Introduction". The notation is adapted from there, in particular references indicated by "A" distinguish the references of that chapter from that of the preceding paper.

Let $R_1 < 1$, and assume that $\max\{I_1, I_2\}^2$ is much smaller than $\min\{M_1, M_2\}$, with $\min\{M_1, M_2\} \to \infty$. Then Theorem A.2 indicates that the epidemic process $\boldsymbol{E}_{\boldsymbol{I}}^{\boldsymbol{M}}$ (A.1) and its approximating branching process $\boldsymbol{Z}_{\boldsymbol{I}}$ (A.2), given that they start with $\boldsymbol{I} = (I_1, I_2)$ infectious and $\boldsymbol{M} = (M_1, M_2)$ susceptibles animals, coincide with high probability.

Hence the results in the preceding paper can be used to approximate the time to extinction for $\boldsymbol{Z}_{\boldsymbol{I}}$, which is at the same time also an approximation for the time to extinction in $\boldsymbol{E}_{\boldsymbol{I}}^{\boldsymbol{M}}(t)$. For $q_i(t)$ $(i = 1, 2)$, system (2) in the preceding paper yields

$$\frac{d\boldsymbol{q}(t)}{dt} = \begin{pmatrix} -\lambda_1 & \rho\theta \\ \lambda_2 & -\lambda_2 \end{pmatrix} \boldsymbol{q}(t) - \begin{pmatrix} \rho\theta q_1(t) q_2(t) \\ 0 \end{pmatrix} =: \boldsymbol{B}\boldsymbol{q}(t) - \boldsymbol{v}(t)\,,$$

where $\boldsymbol{q}(t) = (q_1(t), q_2(t))^T$. Since $\|\boldsymbol{J}\| \leq 2$ for $\boldsymbol{Z}_{\boldsymbol{I}}(0) = (1, 0)^T$ and $\boldsymbol{Z}_{\boldsymbol{I}}(0) = (0, 1)^T$, we can take $\alpha = 1$. Theorem 2.1 in the preceding paper implies that $\boldsymbol{B}$ has a unique real largest eigenvalue $-r$, with corresponding positive left $\boldsymbol{f}_1^T$ and right $\boldsymbol{b}_1$ eigenvectors, which are given by $-r = (-(\lambda_1 + \lambda_2) + \sqrt{D})/2$ with $D = (\lambda_1 + \lambda_2)^2 - 4\lambda_2(\lambda_1 - \rho\theta)$, and the corresponding positive left and right eigenvectors $\boldsymbol{f}_1$ and $\boldsymbol{b}_1$ satisfying $\|\boldsymbol{f}_1\| = 1$ and $|\boldsymbol{f}_1^T \boldsymbol{b}_1| = 1$ are straightforward to construct.

As in the application to the parasitic protozoa *Toxoplasma gondii* in the preceding paper, the approximate time to extinction for the branching process (A.2) is given by $\tilde{T}_{\boldsymbol{I}} = \log C_{\boldsymbol{I}}/r + V/r$, where $C_{\boldsymbol{I}} = c_1 I_1 + c_2 I_2$ with $c_1, c_2 > 0$ constants which can be computed much as before, and V is a Gumbel random variable. This approximation can now be compared to the distribution of the true extinction time T of the epidemic model (A.1). The true distribution is theoretically not amenable, and thus needs to be computed by simulation.

The parameters of the epidemic process (A.1) are chosen such that they reasonably reflect a "typical" situation in Central Asia. The population ratio ϱ is approximated by 10 based on an estimate of 10.368 from (unpublished) field data in Kazakhstan, where during a purgation study in dogs, the owners have been asked how many sheep and dogs they own. It is assumed that there are $n^{(1)} = 500$ dogs, and thus $n^{(2)} = n^{(1)}\rho = 5000$ sheep. The death rate λ_2 is set to 0.5 based on an estimate of 0.491 ($95\% CI : 0.473, 0.501$) in a sheep sample from Kazakhstan (Torgerson et al. 2003b). We have seen in the second paper of this thesis that the loss

of infection rate is about $1 - 1.2$ infections per dog per year, and thus we choose $\lambda_1 \in \{1, 1.2\}$. The contact rate θ is chosen such that $R_1 < 1$. For our application, we take $\theta \in \{0.01, 0.05\}$.

Figure 4 displays the distribution of the approximate time to extinction and the simulated distribution of the true time to extinction for the different parameter settings. The approximate time to extinction is well in line with the simulated distribution of the true time for all settings. Longer mean times to extinction are observed for decreasing values of λ_1 (see in Figure 4, (x1)-(x2) for x=a,b,c), and for increasing values of θ (see in Figure 4, (x2)-(x3) for x=a,b,c). These observations can be explained as follows. Recall the construction argument of the branching process in Section A.3, where the life histories of infections in dogs are specified as (L_1, ξ_1), with L_1 exponentially distributed with mean $1/\lambda_1$ and with ξ_1 a Poisson process of rate $\theta\rho$ at the points of which sheep make infective contacts with its excreta, and the life histories for sheep with (L_2, ξ_2), where L_2 is exponentially distributed with rate λ_2 and $\xi_2[0, L_2) = 0$ and $\xi_2\{L_2\} = 1$, since an infected sheep is connected with exactly one dog and the infection is transmitted at death of the sheep. Let $\mathcal{P}_1$ be a Poisson process with rate λ_1. Let $T_1, T_2, \ldots$ be the arrival times of the Poisson process. Introduce two marked point processes based on $\mathcal{P}_1$. In the first, mark all occurrence times of $\mathcal{P}_1$ with probability 1. In the second, mark the occurrence times with probability $\lambda_1'/\lambda_1 < 1$, where $\lambda_1' < \lambda_1$. Define L_1 as the first marked occurrence time. Hence $L_1 = T_1$ for the first marked process and $L_1 = T_j$ with probability $(1 - \lambda_1'/\lambda_1)^{j-1}(\lambda_1'/\lambda_1)$, $j \geq 1$, for the second. Note that for the second process, L_1 has the exponential distribution with mean $1/\lambda_1'$, and so corresponds to the lifetime of an infection of a dog, when the recovery rate λ_1' is smaller than λ_1. Hence each infection duration can be constructed to be longer almost surely in dogs for the latter process, so that dogs in the second process will infect more sheep if the same constant process ξ_2 is used in both cases. Since infection is transmitted back to the dog population with probability 1, the second process implies an increased time to extinction almost surely, and hence also in mean. A similar argument can be used to show that increasing θ implies increasing the mean time to extinction. Finally, increasing values of the initial conditions I_1 and I_2 imply longer mean times to extinction since $\tilde{T}_{\boldsymbol{I}}$ grows like $\log C_{\boldsymbol{I}} = \log(c_1 I_1 + c_2 I_2)$, with $c_1, c_2 > 0$ fixed. Despite the shift of the mean, it is clear from the definition of $\tilde{T}_{\boldsymbol{I}}$ that the shape remains the same for different values of the initial conditions (see in Figure 4, (ai)-(ci) for i=1,2,3).

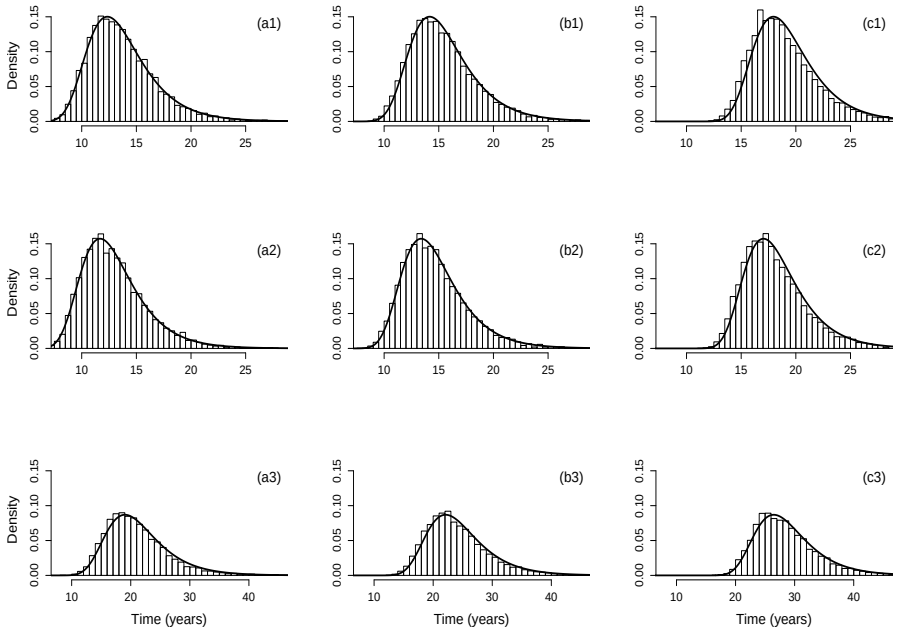

Figure 4: *Density distribution of* $\tilde{T}_{\boldsymbol{I}}$ *(solid line) versus the simulated distribution of the true extinction time (histogram of* 10000 *simulations) for model (A.1), with fixed parameters* $n^{(1)} = 500$, $n^{(2)} = 5000$, $\lambda_2 = 0.5$. *The parameter pair* (λ_1, θ) *is* $(1, 0.01)$ *for* (a1)-(c1), $(1.2, 0.01)$ *for* (a2)-(c2) *and* $(1.2, 0.05)$ *for* (a3)-(c3). *The initial conditions* (I_1, I_2) *are* $(20, 100)$ *for* (a1)-(a3), $(100, 200)$ *for* (b1)-(b3), *and* $(100, 1000)$ *for* (c1)-(c3).

References

Aminzhanov, M. (1975), 'Duration of the life of Echinococcus granulosus in the organism of dogs', *Veterinariia* **12**, 70–72.

Anderson, R. M. (1974), 'Population dynamics of the cestode Caryophyllaeus laticeps in the bream', *J Anim Ecol* **43**, 305–321.

Anderson, R. M. & May, R. M. (1982), *Population biology of infectious diseases*, Berlin: Springer.

Athreya, K. B. & Ney, P. E. (1972), *Branching Processes*, Berlin: Springer.

Ball, F. (1983), 'The threshold behaviour of epidemic models', *J. Appl. Prob.* **20**, 227–241.

Ball, F. & Donnelly, P. (1995), 'Strong approximations for epidemic models', *Stoch. Procs. Applics.* **55**, 1–25.

Balling, T. E. & Pfeiffer, W. (1997), 'Frequency distributions of fish parasites in the perch *Perca fluviatilis* l. from Lake Constancee', *Parasitol Res* **83**, 370–373.

Barbour, A. D. & Kafetzaki, M. (1991), 'Modeling the overdispersion of parasite loads', *Math Biosci* **107**, 249–253.

Bass, F. (1969), 'A new product growth model for consumer durables', *Management Science* **15**, 215–227.

Bliss, C. I. & Fisher, R. A. (1953), 'Fitting the negative binomial distribution to biological data', *Biometrics* **9**, 176–200.

Boag, B., Topham, P. B. & Webster, R. (1989a), 'Spatial distribution on pasture of infective larvae of the gastro-intestinal nematode parasites of sheep', *Int J Parasitol.* **19**, 681–685.

Budke, C. M., Qiu, J., Craig, P. S. & Torgerson, P. R. (2005), 'Modeling the transmission of Echinococcus granulosus and Echinococcus multilocularis in dogs for a high endemic region of the Tibetan plateau', *Int J Parasitol.* **35**, 163–170.

Diekmann, O., Heesterbeek, J. A. & Metz, J. A. (1990), 'On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations', *J Math Biol.* **28**, 365–382.

Eckert, J. & Deplazes, P. (2004), 'Biological, epidemiological and clinical aspects of Echinococcosis, a zoonosis of increasing concern', *Clin Microbiol Rev.* **17**, 107–135.

Eckert, J., T., F. K., Zahner, H. & Deplazes, P. (2005), *Lehrbuch der Parasitologie für die Tiermedizin*, 1 edn, Stuttgart:Enke Verlag.

Economides, P. & Cristofi, G. (2002), *Cestode zoonoses: Echinococcosis and cysticercosis. An emergent and global problem*, 3 edn, NATO Science Series:IOS Press Amsterdam.

Flütsch, F., Heinzmann, D., Mathis, A., Hertzberg, H., Stephan, R. & Deplazes, P. (2008), 'Case-control study to identify risk factors for bovine cysticercosis on farms in Switzerland', *Parasitology* **135**, 641–646.

Gemmell, M. A., Lawson, J. R. & Roberts, M. G. (1986), 'Population dynamics in echinococcosis and cysticercosis: biological parameters of Echinococcus granulosus in dogs and sheep', *Parasitology* **92**, 599–620.

Harris, T. E. (1963), *The theory of branching processes*, Berlin: Springer.

Heesterbeek, J. A. & Roberts, M. G. (2007), 'The type-reproduction number T in models for infectious disease control', *Math. Biosci.* **206**, 3–10.

Heinzmann, D. & Torgerson, P. R. (2008), 'Evaluating parasite densities and estimation of parameters in transmission systems', *Parasite* **15**, 477–483.

Herbert, J. & Isham, V. (2000), 'Stochastic host-parasite interaction models', *J. Math. Biol.* **40**, 343–371.

Jagers, P. (1975), *Branching Processes with Biological Applications*, New York: Wiley and Sons.

Jagers, P., Klebaner, F. C. & Sagitov, S. (2007), 'On the path to extinction', *PNAS* **104**, 6107–6111.

Lloyd-Smith, J. O. (2007), 'Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases', *PLoS ONE* **2**, doi:10.1371/journal.pone.0000180.

Luchsinger, C. J. (2001), 'Stochastic models of a parasitic infection, exhibiting three basic reproduction ratios', *J Math Biol* **42**(6), 532–554.

Permin, A. & Hansen, J. W. (1994), 'Review of echinococcosis/hydatidosis: a zoonotic parasitic disease', *World Animal Review (FAO)* **78**.

Pugliese, A., Rosa, R. & Damaggio, M. L. (1998), 'Analysis of a model for macroparasitic infection with variable aggregation and clumped infections', *J Math Biol* **36**, 419–447.

Rapsch, C., Dahinden, T., Heinzmann, D., Torgerson, P. R., Braun, B., Deplazes, P., Hurni, L., Bär, H. & Knubben-Schweizer, G. (2008), 'An interactive map to assess the potential spread of lymnaea truncatula and the free-living stages of fasciola hepatica in switzerland', *Vet Parasitol.* **154**, 242–249.

Rüegg, S., Heinzmann, D., Barbour, A. D. & Torgerson, P. R. (2008), 'Estimating the transmission dynamics of theileria equi and babesia caballi in horses', *Parasitology* **135**, 555–565.

Roberts, M. G., Lawson, J. R. & Gemmell, M. A. (1986), 'Population dynamics in echinococcosis and cysticercosis: Mathematical model of the life-cycle of Echninococcus granulosus', *Parasitology* **92**, 621–641.

Sewastjanow, B. A. (1974), *Verzweigungsprozesse*, Berlin: Akademie-Verlag.

Tallis, G. M. & Leyton, M. (1969), 'Stochastic models of populations of helminthic parasites in the definitive host', *Math Biosci* **4**, 39–48.

Tanner, C. E., Curtis, M. A., Sole, T. D. & K., G. (1980), 'The nonrandom, negative binomial distribution of experimental trichinellosis in rabbits', *J. Parasitol.* **66**, 802–805.

Torgerson, P. R., Burtisurnov, K. K., Shaikenov, B. S., Rysmukhambetova, A. T., Abdybekova, A. M. & Ussenbayev, A. E. (2003a), 'Modelling the transmission dynamics of Echinococcus granulosus in dogs in rural Kazakhstan', *Parasitology* **126**, 417–424.

Torgerson, P. R., Karaeva, R. R., Corkeri, N., Abdyjaparov, T. A., Kuttubaev, O. T. & Shaikenov, B. S. (2003c), 'Cystic echinococcosis in humans in Kyrgystan: an epidemiological study', *Acta Tropica* **85**, 51–61.

Torgerson, P. R., Oguljahan, B., Muminov, M. E., Karaeva, R. R., Kuttubaev, O. T., Aminjanov, M. & Shaikenov, B. (2006), 'Present situation of cystic echinococcosis in Central Asia', *Parasitol Int.* **55**, 207–212.

Torgerson, P. R., Shaikenov, B. S., Rysmukhambetova, A. T., Ussenbayev, A. E., Abdybekova, A. M. & Burtisurnov, K. K. (2003b), 'Modelling the transmission dynamics of Echinococcus granulosus in sheep and cattle in Kazakhstan', *Vet Parasitol* **114**, 143–153.

Torgerson, P. R., Williams, D. H. & Abo-Shehada, M. N. (1998), 'Modelling the prevalence of Echinococcus and Taenia species in small ruminants of different ages in northern Jordan', *Vet Parasitol* **79**, 35–51.

Woolhouse, M. E. J., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L. K., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K. & Anderson, R. M. (1997), 'Heterogeneities in the transmission of infectious agents: Implications for the design of control programs', *PNAS* **94**, 338–342.

Zhang, Z. Q., Chen, P. R., Wang, K. & Wang, X. Y. (2008), 'Overdispersion of Allothrombium pulvinum larvae (Acari: Trombidiidae) parasitic on Aphis gossypii (Homoptera: Aphididae) in cotton fields', *Ecol Entomology* **18**, 379–384.